

Detection and Prevention of Cyberbullying on Social Media

SEMIU DOLAPO SALAWU
Doctor of Philosophy (by Research)

ASTON UNIVERSITY
January 2021

© Semiu Dolapo Salawu, 2021

Semiu Dolapo Salawu asserts his moral rights to be identified as the author of this thesis

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without appropriate permission or acknowledgement.

Aston University

Detection and Prevention of Cyberbullying on Social Media

Semiu Dolapo Salawu

Doctor of Philosophy, 2021

Thesis Summary

The Internet and social media have undoubtedly improved our abilities to keep in touch with friends and loved ones. Additionally, it has opened up new avenues for journalism, activism, commerce and entertainment. The unbridled ubiquity of social media is, however, not without negative consequences and one such effect is the increased prevalence of cyberbullying and online abuse. While cyberbullying was previously restricted to electronic mail, online forums and text messages, social media has propelled it across the breadth of the Internet, establishing it as one of the main dangers associated with online interactions. Recent advances in deep learning algorithms have progressed the state of the art in natural language processing considerably, and it is now possible to develop Machine Learning (ML) models with an in-depth understanding of written language and utilise them to detect cyberbullying and online abuse. Despite these advances, there is a conspicuous lack of real-world applications for cyberbullying detection and prevention. Scalability; responsiveness; obsolescence; and acceptability are challenges that researchers must overcome to develop robust cyberbullying detection and prevention systems.

This research addressed these challenges by developing a novel mobile-based application system for the detection and prevention of cyberbullying and online abuse. The application mitigates obsolescence by using different ML models in a “plug and play” manner, thus providing a mean to incorporate future classifiers. It uses ground truth provided by the end-user to create a personalised ML model for each user. A new large-scale cyberbullying dataset of over 62K tweets annotated using a taxonomy of different cyberbullying types was created to facilitate the training of the ML models. Additionally, the design incorporated facilities to initiate appropriate actions on behalf of the user when cyberbullying events are detected.

To improve the app’s acceptability to the target audience, user-centred design methods were used to discover stakeholders’ requirements and collaboratively design the mobile app with young people. Overall, the research showed that (a) the cyberbullying dataset sufficiently captures different forms of online abuse to allow the detection of cyberbullying and online abuse; (b) the developed cyberbullying prevention application is highly scalable and responsive and can cope with the demands of modern social media platforms (b) the use of user-centred and participatory design approaches improved the app’s acceptability amongst the target audience.

Keywords: cyberbullying detection, cyberbullying prevention, deep learning, participatory design, Twitter, mobile application.

*Dedicated to Habi, Navyd and Taaliah.
The three points of my North Star.
Without you three, I am adrift.
You anchor me when I am rudderless.
To my Mom, thanks for always being there.
And of course Dad, you always wanted me to do this.
I know you are watching from above.
My love for you all is eternal.*

Acknowledgements

Great is the art of beginning, but greater is the art of ending.

This Ph.D research would not have been possible without the help and support of some truly wonderful people. First I would like to thank my supervisor, Dr. Jo Lumsden. For her dedication and incredible support. For holding me up to her incredibly high standards and giving me the push I needed to get me going the many times I got stuck. For knowing when a simple late night email or a quick Skype call would give me the motivation I needed to push through. Thank you for your detailed reviews and comments, I will never look at a semicolon in the same way again. I feel truly blessed to have you as my supervisor.

Special thanks to my associate supervisor, Yulan He, for her invaluable input and technical guidance as I navigated the wonderful world of natural language processing and machine learning. For gently nudging me in the right direction when I was running down blind alleys.

A big thank you to the West Midlands Police and all my participants. To the users of BullStop that contacted me to show their appreciation and provide invaluable feedback, you gave more motivation than you could have ever imagined.

Most importantly, to Habi, my Jewel of Inestimable Value, thank you for being the most supportive wife I could have ever asked. And finally, to Navyd and Taaliah, yes, you can have daddy back now.

Publications Arising from this Thesis

1. Salawu, S., He, Y., and Lumsden, J. (2017). Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing*, 11(1), 3-24.

This paper provides an in-depth survey of approaches to the automated detection of cyberbullying and online abuse.

2. Salawu, S., He, Y., and Lumsden, J. (2020). BullStop: A Mobile App for Cyberbullying Prevention. In *Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations* (pp. 70-74).

This paper provides an overview of the BullStop application and the experiments conducted to identify the best performing machine learning model to use in the application.

3. Salawu, S, Lumsden, J. and He, Y. (2020). A Large-Scale Multi-Label Twitter Dataset for Online Abuse Detection. Accepted for publication at The 5th Workshop on Online Abuse and Harms (WOAH 2021) of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021).

This paper discusses the cyberbullying dataset created as part of this research program.

4. Salawu, S, Lumsden, J. and He, Y. (2020). A Responsive and Scalable Mobile-Based System for Cyberbullying Detection and Prevention. Accepted for publication by the International Journal of Bullying Prevention's Special Issue on The Use of Artificial Intelligence to Address Online Bullying and Abuse.

This is an extended version of the above paper entitled "BullStop: A Mobile App for Cyberbullying Prevention" and provides an in-depth discussion of the application system and its development.

Contents

Acknowledgements	4
Publications Arising from this Thesis	5
List of Figures	11
List of Tables	14
1 Introduction	15
1.1 Research Motivation	15
1.2 Research Question and Objectives	16
1.3 Contribution to Scientific Knowledge	19
1.4 Thesis Structure	21
2 Literature Review	23
2.1 Introduction	23
2.2 Definition, Prevalence and Consequences of Cyberbullying	24
2.3 Cyberbullying Detection Techniques	29
2.3.1 Traditional Machine Learning Approaches	44
2.3.2 Deep Learning Approaches	50
2.3.3 Transformer-Based Cyberbullying Detection	54
2.4 Systems, Tools and Applications for Preventing Cyberbullying	59
2.5 Mobile Apps for Cyberbullying Prevention and Mitigation	64
2.5.1 Method	65
2.6 Cyberbullying Prevention by Social Media Platforms	77
2.7 Summary	78
3 Fine-Grained Detection of Cyberbullying on Social Media	81
3.1 Introduction	81
3.2 Existing Cyberbullying Datasets	82
3.3 A Twitter-Based Dataset for Detecting Cyberbullying and Offensive Language	86
3.3.1 Dataset Objective	86
3.3.2 Labels	86
3.3.3 Data Collection	89
3.3.4 Annotation Process	90
3.3.5 Preprocessing	91
3.3.6 Dataset Analysis	92
3.3.7 Bias and Practical Usage Implication	94
3.3.8 Dataset Availability	96

3.4	Best Performing Model Selection	96
3.4.1	Models Evaluation and Best Performing Model Selection	96
3.4.2	Evaluating the Dataset's Fitness for Purpose	100
3.5	Summary	102
4	Stakeholders' Perspectives on Cyberbullying Prevention	104
4.1	Introduction	104
4.2	Adult Stakeholders' Perspectives on Cyberbullying Prevention	105
4.2.1	Focus Group and Participant Recruitment	106
4.2.2	Content Analysis	111
4.2.3	Emergent Themes	111
4.2.3.1	Concerns About Cyberbullying	113
4.2.3.2	Current Strategies and Solutions are in Need of Improvement	115
4.2.3.3	Encouraging Positive Behaviours and Online Safeguarding are Key Features for the Proposed App	117
4.2.3.4	Report and Block Online Abusers	118
4.2.4	Discussion	118
4.2.5	Study Limitations	121
4.3	Understanding Young People's Attitudes to Cyberbullying and its Prevention	122
4.3.1	Pre-Study Questionnaire	123
4.3.2	Interviewees	128
4.3.3	Emergent Themes	129
4.3.3.1	Cyberbullying Occurrence Intensifies in Early Teenhood and Extends into the Late Teens	129
4.3.3.2	Appearance and Identity are Common Bullying Themes . .	132
4.3.3.3	Cyberbullying on Facebook and Twitter is More Public Compared to Snapchat and Instagram, Where it is More Personal and Targeted	133
4.3.3.4	Cyberbullies are Often Known to Their Victims	135
4.3.3.5	Fear of Reprisals and Inadequate Responses Discourages Cyberbullying Reporting	137
4.3.3.6	School Should Intensify Cyberbullying Prevention Efforts .	139
4.3.3.7	Relevant Advice and Punitive Actions are the Critical Features for the Proposed App	140
4.3.3.8	Young People Would Rather Report Cyberbullying Anonymously than get Directly Involved	142
4.3.4	Discussion	143
4.3.5	Study Limitations	147
4.4	Summary	148
5	Participatory Design of the BullStop Mobile Application	150
5.1	Introduction	150
5.2	Participatory Design	151
5.3	Study Design	152
5.4	Recruitment of Participants	154
5.5	Design and Prototyping	155
5.5.1	Design Conceptualisation	156
5.5.2	Low Fidelity Prototyping	159

5.5.3	Findings from the First Participatory Design Session	163
5.5.3.1	Use a Cool and Symbolic Logo	163
5.5.3.2	The App's Name Should be Short and Catchy	164
5.5.3.3	Use a Neutral but Friendly and Welcoming Colour like Blue	164
5.5.3.4	The App Should Not Look Childish or be Patronising to the Young Target Audience	165
5.5.4	High-Fidelity Prototyping	165
5.5.5	Findings from the Second Participatory Design Session	172
5.5.5.1	Provide Shortcuts to Key Components to Facilitate First-time Use	173
5.5.5.2	Use familiar Icons to Signpost Actions	173
5.5.5.3	Reassure Users that the App is Secure	174
5.5.5.4	Emphasise that the App Supports Multiple Social Media Platforms	174
5.5.5.5	Creating a Comprehensive Online Presence for the App can Reassure and Attract Potential Users	175
5.6	Reflections on Participatory Design with Young People	175
5.6.1	Self-pride	177
5.6.2	Learning	178
5.6.3	Empowerment	178
5.7	Study Limitations	179
5.8	Summary	180
6	The BullStop Mobile Application	182
6.1	Introduction	182
6.2	Application Features Implementation	183
6.3	Overview of the BullStop Mobile Application	185
6.4	Android Application Framework	187
6.5	User Interface	189
6.5.1	End User Licence Agreement	189
6.5.2	Forgot Password	189
6.5.3	Sign In	190
6.5.4	Sign Up	191
6.5.5	Home	192
6.5.6	Social Media Login	192
6.5.7	Enable BullStop	192
6.5.8	Sent Messages	193
6.5.9	Received Messages	193
6.5.10	Deleted Messages	195
6.5.11	Message Checker	196
6.5.12	Manage Contacts	196
6.5.13	Report	197
6.5.14	Help	197
6.5.15	Helplines	198
6.5.16	Settings	198
6.5.17	Message Settings	198
6.5.18	Detection Settings	198

6.5.19	Social Account Settings	199
6.5.20	Tour	199
6.6	Application Logic	199
6.7	SQLite Database	200
6.8	Cloud Backend	200
6.8.1	Webhook	202
6.8.2	Message Queue	203
6.8.3	Abuse Detection Module (ADM)	203
6.8.4	Model Repository	205
6.8.5	Online Training Module	205
6.8.6	Remote Database	206
6.8.7	Synchroniser	206
6.8.8	Real-Time API	206
6.8.9	Marshaller	207
6.9	Limitations and Challenges	207
6.9.1	Summary	211
7	Evaluation of the BullStop Mobile Application	213
7.1	Introduction	213
7.2	Evaluation of the System's Responsiveness	214
7.3	Evaluation of the System's Scalability	215
7.4	Mitigating System's Obsolescence	216
7.5	'Lab'-Based Evaluation of the Application's Acceptability to Users	217
7.5.1	Study Design	217
7.5.2	Participant Recruitment	218
7.5.3	Findings and Discussion	218
7.5.3.1	Easy to Use	219
7.5.3.2	Well Designed	220
7.5.3.3	Appropriate Branding	221
7.5.3.4	Good Overall Performance	222
7.5.3.5	User Empowerment	222
7.5.3.6	Reflective and Educative	224
7.5.3.7	Useful and Unique	225
7.5.3.8	Suggested Improvements	226
7.6	Field-Based Evaluation of the Application's Acceptability	228
7.6.1	Study Design	228
7.6.2	Online Presence	231
7.6.3	Analysis of Users' Engagement with the Mobile App	232
7.6.4	Analysis of Users' Gender and Age Demography	235
7.6.5	Analysis of Application Usage Patterns	236
7.6.6	Findings and Discussion	239
7.6.6.1	App Usage by Young People	239
7.6.6.2	Perceived Usefulness of the App	241
7.6.6.3	Perceived Usability of the App	244
7.6.6.4	Perceived Responsiveness of the App	245
7.6.6.5	Users' Favourite Aspects of the App	245
7.6.6.6	Suggested Areas for Improvement	246

7.7	Summary	247
8	Reflections on the Design Approach	249
8.1	Introduction	249
8.2	Learning Outcomes	249
8.2.1	Reflection #1: Inclusion and Representation are Vital	250
8.2.2	Reflection #2: Build Rapport with Participants from the Onset Via Frequent Communication	251
8.2.3	Reflection #3: When Soliciting Knowledge from Participants, Be Prepared to Introduce Procedural Flexibility to Progress	252
8.2.4	Reflection #4: Maximise Opportunities to Engage with Participants .	253
8.2.5	Reflection #5: Participants Empowerment Encourages Active Participation	254
8.3	Summary	254
9	Conclusions, Contribution to Knowledge, and Further Research	256
9.1	Thesis Conclusion	256
9.2	Contributions to Scientific Knowledge	261
9.3	Future Research	263
	References	267

List of Figures

2.1	Neural Network vs Deep Neural Network.	51
2.2	The Transformer Model Architecture. Source: Vaswani et al., 2017.	55
2.3	Use of reflective interface on Instagram, Twitter and Facebook to warn users of sensitive content.	60
2.4	Example of a tweet identified as glorifying violence by Twitter.	61
2.5	Mobile apps survey search and selection process flow chart.	66
2.6	'Similar Apps' listing on an app's page on the Google Play Store.	67
2.7	'You May Also Like' apps listing on an app's page on the Apple App Store.	67
2.8	BBC Own It sample screens	74
2.9	ReThink sample screens	75
2.10	Bark For Kids sample screens	75
2.11	Bosco Family Safety sample screens	76
2.12	Sentry Parental Control sample screens	76
2.13	Surfie Parental Control sample screens	77
2.14	MMGuardian Parental Control sample screens	77
3.1	Distribution of tweet counts and number of labels assigned.	93
3.2	Distribution of tweet counts and number of labels assigned.	94
3.3	Example of ROC Curve.	97
4.1	Social media platforms used by respondents.	124
4.2	Frequency of social media interaction.	124
4.3	Means of accessing social media.	125
4.4	Number of participants that have been cyberbullied.	125
4.5	Period when the cyberbullying incident occurred.	126
4.6	Type of online abuse experienced.	126
4.7	Respondents' response to cyberbullying when they are the victim.	127
4.8	Number of respondents that have witnessed a friend being cyberbullied.	127
4.9	Respondents' response to cyberbullying when a friend is the victim.	127
5.1	Overview of the design process.	153
5.2	Room arrangement showing work surface and breakout areas.	156
5.3	Participants working collaboratively and writing down ideas.	157
5.4	Spider diagram of key features created during the first PD session.	158
5.5	Essential features shortlist compiled during the first PD session.	159
5.6	Home screen prototypes from created using Pair Design.	161
5.7	Home screen base prototype.	162
5.8	Home screen base prototype.	162

5.9 Deleted Messages screen prototype.	163
5.10 Spider diagram of key features for the proposed mobile apps.	166
5.11 The logo for the Proposed mobile app.	168
5.12 Splash Prototype.	168
5.13 EULA Prototype.	168
5.14 Privacy Prototype.	168
5.15 Account Prototype	169
5.16 Home Prototype.	169
5.17 Setting Prototype.	169
5.18 Message Settings Prototype.	170
5.19 Detection Settings Prototype.	170
5.20 Social Account Settings Prototype.	170
5.21 Received Prototype.	170
5.22 Sent Prototype.	170
5.23 Deleted Prototype.	170
5.24 Message Review Prototype.	171
5.25 Message Checker Prototype.	171
5.26 Manage Contacts Prototype.	171
5.27 Help Prototype.	172
5.28 Tour Prototype.	172
5.29 Helplines Prototype.	172
5.30 Error Message Prototype	172
5.31 Suggested shortcuts	173
6.1 Logical overview of the BullStop Application System.	186
6.2 Key components of the BullStop System and their functions.	186
6.3 Android Platform architecture. Source: developer.android.com.	188
6.4 Navigational map of the app.	190
6.5 End User Licence Agreement Screen.	191
6.6 Forgot Password Screen.	191
6.7 Sign Up Screen.	192
6.8 Sign In Screen.	192
6.9 Home Screen.	193
6.10 Login Screen.	193
6.11 Enable BullStop Dialog.	194
6.12 Sent Messages Screen.	194
6.13 Message Review Screen.	194
6.14 Received Messages Screen.	195
6.15 Deleted Messages Screen.	195
6.16 Message Checker Screen.	197
6.17 Manage Contacts Screen.	197
6.18 Report Screen.	197
6.19 Help Screen.	197
6.20 Helplines Screen.	198
6.21 Settings Screen.	198
6.22 Message Settings Screen.	199
6.23 Detection Settings Screen.	199

6.24 Social Account Settings Screen.	199
6.25 Tour Screens	200
6.26 High Level Architecture of the Cloud Backend	202
6.27 Android device categorisation based on screen size and pixel density.	208
7.1 Graph illustrating Message Consumption Rate under artificially-induced load.	216
7.2 Sent Messages and Received Messages icons.	227
7.3 Sent Messages and Received Messages icons.	229
7.4 Online questionnaire welcome page.	229
7.5 Google Play Console page for BullStop.	230
7.6 Firebase Dashboard Play for BullStop.	230
7.7 BullStop listing on the Google Play Store.	231
7.8 BullStop app website.	232
7.9 Traffic to the BullStop website during the evaluation period.	232
7.10 Google search results for 'bullstop'.	233
7.11 Total number of active, lost and acquired users and visitors.	233
7.12 Users acquisition and loss before and after the press release.	234
7.13 How the app was discovered on the app store by users.	235
7.14 Age groups and gender of app users.	236
7.15 Age distribution of Twitter users.	236
7.16 The Message Checker screen's check button.	238
7.17 Age groups and gender of respondents.	240
7.18 Social media platforms usage amongst BullStop users.	241
7.19 How BullStop users found out about the app.	241
7.20 Users overall rating for the app.	244
7.21 Sent Users responsiveness rating for the app.	245
9.1 Sample email messages from BullStop users	264

List of Tables

2.1	Discovered literature on cyberbullying detection and the approaches used.	43
2.2	Automated cyberbullying prevention studies reviewed.	64
2.3	Overview of dummy accounts used to conduct the tests.	70
2.4	Description of tests conducted to evaluate the apps.	71
2.5	Results of evaluation for cyberbullying prevention apps.	72
3.1	Comparison of cyberbullying and offensive content datasets.	87
3.2	Annotation scheme with examples.	88
3.3	Total number of tweets each label was assigned to.	92
3.4	Example row from the dataset.	96
3.5	Results of classification experiments.	99
3.6	Results of cross-domain experiments.	101
4.1	Thematic Analysis Phases and Tasks.	112
4.2	Adult stakeholder's desired features for the proposed app.	120
4.3	Summary of interviewees' ethnic groups.	128
4.4	Young People's desired features for the proposed app.	147
5.1	Summary of PD participants profiles.	155
5.2	Features prioritisation using dot voting.	158
5.3	Re-prioritised features.	159
5.4	Pair Design teams' composition.	161
6.1	Overview of suggested application features and their implementation status.	185
6.2	Weights for offensive labels.	205
7.1	Results of Responsiveness Evaluation Experiments.	214
7.2	Overview of evaluation study participants	219
7.3	Top Ten screens users spent the most time on.	238

Chapter 1: Introduction

1.1 Research Motivation

On August 2nd, 2013, Jo Smith discovered the dead body of her 14-year-old younger sister, Hannah, in their family home in Lutterworth, Leicestershire, UK. According to her family, Hannah suffered extensive abuse online on the social website Ask.fm, including several anonymous messages encouraging her to commit suicide (BBC News, 2013). Ask.fm is a social media website that allows users to post questions and solicit responses; respondents can choose to be anonymous when replying to a question. A month later in Florida, USA, 12-year-old Rebecca Ann Sedwick jumped to her death after being repeatedly harassed online by a group of similarly-aged girls. According to police investigators, Rebecca's tormentors used various means to harass her, including hacking her Facebook account, sending abusive messages via text and Instant Messaging (IM) platforms and even following her across various social networking sites posting hateful comments about her (Alvarez, 2013). Brandy Vela, an 18-year-old student, shot herself dead in front of her parents at their home in Texas, USA, following years of relentless online abuse, including fake profiles created using her pictures offering 'sex for free' (Pasha-Robinson, 2012).

Unfortunately, incidents such as these have become part of the ongoing narrative about cyberbullying, and it is estimated that as much as 59% of young people will have experienced some form of cyberbullying by the time they become young adults (Pew Research Center, 2018). Nowhere is cyberbullying more prevalent than on social media, where 69% of reported cyberbullying incidents occurred (Ofcom Research, 2019). In recent years, reports of cyberbullying-related suicides have increased in the media (Hinduja and Patchin, 2010) and, unsurprisingly, this has focussed significant attention on cyberbullying and its prevention. Cyberbullying has been linked to mental health issues

such as low self-esteem, depression, anxiety and substance abuse (Hinduja and Patchin, 2010; Khine *et al.*, 2020; Martínez-Monteagudo *et al.*, 2020) and the pervasiveness of social media that has allowed online abuse to extend beyond geographic boundaries has introduced additional complications in the efforts to prevent it.

Social media platforms like Twitter, Facebook, and Instagram have responded to this increased proliferation of online abuse by introducing policies and features to combat and mitigate cyberbullying and its effects. These have included preventing the creation of multiple accounts using similar details and frequent suspension of abusive users. Online social networks also employ human moderators to review thousands of posts daily for inappropriate content. Unfortunately, despite these efforts, cyberbullying and online harassment remain a significant online risk for many young people. Furthermore, as cyberbullying is highly subjective, what is deemed offensive differs amongst people and human moderators can only apply common-sense rules in making a judgement.

Research efforts on cyberbullying detection have been primarily focused on developing algorithms for detecting cyberbullying and offensive language automatically. While this has contributed to advancing the state-of-the-art in the automated detection of cyberbullying, it has largely overlooked the development of novel, viable tools that can use these state-of-the-art algorithms to tackle online abuse in the real world. Researchers face four key challenges in developing cyberbullying detection systems: scalability; responsiveness (Rafiq *et al.*, 2018); obsolescence; and acceptability.

Research investigating how these challenges can be resolved is rare and is certainly an area in need of attention to improve the availability of automated online abuse prevention solutions. This research programme, therefore, focuses on addressing these challenges as outlined in the following sections.

1.2 Research Question and Objectives

Scalability is the ability of a system to increase its capacity by expanding the quantity of lower-level resources (e.g., memory, CPU) consumed (Lehrig *et al.*, 2015) while responsiveness is defined as a system's ability to complete a set of tasks within an allotted time (Sweet *et al.*, 2001). Both are properties expected of performant systems and are often not considered by researchers when designing cyberbullying detection

systems (Rafiq *et al.*, 2018; Zois *et al.*, 2018). Subsequently, developed systems to date are unable to cope with the large user volume and scalability demands of modern social media platforms or respond promptly to enable real-time communications (Yao *et al.*, 2019).

Existing cyberbullying prevention systems are primarily focused on improving classifier accuracy (Rafiq *et al.*, 2020), and their implementation has typically involved a tight coupling between the classifier and the encompassing system. This introduces obsolescence into a system's design as the classifier cannot be easily changed (when better machine learning models become available), often necessitating a redesign of the entire system and consigning a limited lifespan to the system from inception. Finally, acceptance by end-users is critical for a system's success and, as many online abuse detection systems are often developed in isolation without consultation with potential end-users (Dinakar *et al.*, 2012), they struggle to gain acceptance amongst the intended audience due to not meeting their expectations (Ashktorab and Vitak, 2016). Only by resolving these concerns will viable automated cyberbullying prevention tools become not only widely available to the public but also widely accepted by the public.

Creating a viable cyberbullying detection and prevention system that addresses the four key challenges of responsiveness, scalability, obsolescence and acceptability as identified above requires a multidisciplinary approach that marries core HCI principles for designing user-focused applications using collaborative design methods with the development of scalable systems that can dynamically utilise various machine learning models for the detection of different types of cyberbullying and offensive language.

This is a non-trivial task, the difficulty of which is highlighted by the lack of such systems despite significant research efforts invested in the automated detection and prevention of cyberbullying and online abuse in recent years.

Within the context of the issues highlighted above, the main research question investigated by this research is:

How can cyberbullying and online abuse be detected and prevented on social media platforms such that the key challenges of scalability, responsiveness, obsolescence and acceptability are adequately addressed?

Two further research questions posed specific to the acceptability challenges are:

1. ***What are stakeholders' needs and expectations for a cyberbullying prevention application and does the use of user-centred and participatory design approaches to design and develop the application result in an application that is an accurate reflection of the stakeholders' expectations as measured by their perception of the tool's usability and usefulness?***
2. ***What constitutes effective practice for engaging stakeholders in user-centred research for the design and development of the cyberbullying prevention mobile application?***

The research reported in this dissertation was aimed at achieving the following objectives:

- (i) understand what cyberbullying prevention tools are currently available and how effective these are;
- (ii) generate a large-scale cyberbullying dataset that can be used to train machine learning models to robustly detect different types of cyberbullying on social media;
- (iii) develop a responsive and scalable cyberbullying detection system that utilises a framework that allows the use of the classifiers trained using the dataset sourced in (ii) and other ML models in a 'plug and play' manner (thus mitigating system obsolescence) to detect cyberbullying instances on social media;
- (iv) incorporate a mechanism to utilise ground truth provided by end-users to create personalised cyberbullying detection classifiers for users;
- (v) apply a user-centred design approach to gain insight into stakeholders' opinions on cyberbullying and its prevention and in so doing understand their requirements for the proposed mobile-based cyberbullying prevention application;
- (vi) use participatory design to collaboratively design the mobile app with young people, creating a design prototype in the process;
- (vii) implement the design prototype as an Android application that is made available to the public via the Google Play Store as a free-of-charge cyberbullying prevention mobile app; and
- (viii) conduct evaluation studies to assess the system's performance in terms of its responsiveness, scalability, and acceptability

1.3 Contribution to Scientific Knowledge

In answering the above questions, this research has made significant contributions to scientific knowledge. These are summarised below and discussed in more depth in Chapter 9.

Conducting a survey that investigated the availability of real-world cyberbullying prevention mobile applications and the effectiveness of the discovered mobile apps.

This highlighted the dearth of practical applications that can be used to aid cyberbullying detection and prevention and provides a starting point for future researchers aiming to address this lack of practical online abuse prevention tools. This aspect of the research is discussed in Chapter 2.

Creation of a new large-scale English multi-label dataset that contains annotated instances of various forms of online abuse and cyberbullying and a significant proportion of offensive content to facilitate the training of cyberbullying and foul language classifiers. The dataset captures different types of online abuse, including less frequent examples of cyberbullying like the use of sarcasm to ridicule, social exclusion and threats that are not typically present in existing datasets. Furthermore, the proportion of offensive content in the dataset (over 80%) is more than that of existing datasets allowing for the training of classifiers without the need for oversampling methods to improve the distribution of offensive samples in the dataset. The dataset's generalisability was validated by cross-domain experiments conducted with two other popular cyberbullying datasets. This dataset expands knowledge and available resources for other researchers in this domain; it is now available to the scientific community at large to further research in this area. The discussion on the activities undertaken to create this dataset is presented in Chapter 3.

Conducting experiments with various traditional machine learning and deep-learning models

trained using the created dataset to identify different forms of online abuse on social media. The investigations led to the selection of the pretrained RoBERTa model as the best performing model and also demonstrated the created dataset's usefulness as a resource for training classifiers to identify online abuse on online social networks. This provided a repeatable process that can be used to evaluate classifiers and datasets for various NLP tasks as outlined in Chapter 3.

Engagement with stakeholders such as young people, parents, clinicians and law enforcement to understand their opinions on cyberbullying and its prevention. This study is the first to engage all the identified stakeholder groups to devise automated strategies for the mitigation and prevention of cyberbullying and online abuse. These strategies were subsequently implemented in the cyberbullying prevention mobile app. The qualitative and quantitative data derived from this engagement reinforced and extended existing scientific knowledge on cyberbullying and the effectiveness of existing prevention strategies and proposed new strategies to aid cyberbullying mitigation and prevention. Crucially, it identified essential features for the cyberbullying prevention application to implement to increase its acceptance by the stakeholders. A detailed account of this phase of the research program is presented in Chapter 4.

The design and implementation of a novel cyberbullying detection and prevention application (Bullstop), which comprises:

- a highly scalable and responsive cloud backend that is capable of utilising different ML models and can generate personalised online abuse detecting classifiers for end-users; and
- an Android application designed collaboratively with young people using a participatory design approach to ensure that the final product is representative of the target audience's needs and is acceptable to them.

This novel application has been validated by target stakeholders and is (a) available for societal good and (b) contributes technical advances and know-how for other researchers working in this and related fields. The findings from this could be used by others to bring cyberbullying detection algorithms into real-world use. Chapters 5 and 6 are focussed on these areas of the research conducted.

The development of an approach for incorporating ethics by design into the development of Artificial Intelligence (AI)-based systems by continuously assessing the human impact of technical decisions made during the development of BullStop. There has been ongoing debates about the ethicality of AI use in everyday application and now more than ever, it is critical that AI-based applications are not only evaluated in terms of their technical performance but also their impact on the society.

Conducting a multi-dimensional evaluation study of BullStop to assess its responsiveness, scalability, and acceptability. The computer-based experiments and the human-based evaluation studies conducted to evaluate the system's performance provide a set of repeatable processes that future researchers can adopt and extend to perform similar evaluation exercises to assess a system's fitness for purpose along both technological and human dimensions. These evaluation activities and their output are discussed in Chapter 7.

Methodological reflections and recommendations from the use of UCD techniques to engage with stakeholders in this domain. This methodological know-how (presented in Chapter 8) will assist future researchers considering similar endeavours.

1.4 Thesis Structure

Chapter 2 presents a review of relevant scientific literature on cyberbullying and its automated detection. It begins with an overview of the pertinent issues relating to cyberbullying, its prevalence on social media and the negative consequences attributed to its perpetuation. This is followed by a discussion of cyberbullying detection approaches that explores the various techniques in use, including traditional machine learning- and deep learning-based methods. Research focussed on developing automated cyberbullying prevention solutions is reviewed, and this is followed by a detailed overview of a novel survey conducted to scope existing commercial and free of charge mobile applications for cyberbullying prevention and social media platform-led initiatives for tackling online abuse.

Chapter 3 describes the creation of the large-scale multi-labelled cyberbullying dataset that was generated to facilitate the training of the Machine Learning (ML) models used in the cyberbullying detection system. It reports on the classification experiments conducted to select the best performing model for use in the application and the results of the cross-domain experiments performed to demonstrate the dataset's generalisability.

Chapter 4 reports on qualitative and quantitative research activities conducted to understand stakeholders' opinions on cyberbullying prevention strategies and how these could and should be best implemented in the mobile application. The output of these activities established the user requirements for the cyberbullying prevention mobile

application developed as part of the research programme, shedding light on previously unexplored issues.

Chapter 5 discusses the participatory design approach adopted to create a prototype for the proposed mobile application by working collaboratively with a group of young people as co-designers. An evaluation of the process in the form of participants' reflections is also presented.

Chapter 6 introduces the BullStop application, its key components and functions. It describes the application screens and sub-systems and discusses the technical challenges encountered when developing the application and the limitations introduced by these challenges.

Chapter 7 reports on the computer-based experiments conducted to evaluate the system's performance in terms of its responsiveness and scalability, and the human-based evaluation studies conducted to investigate the application's perceived usefulness and usability. It also reports on an analysis of the application usage data recorded by the mobile application over the course of a field trial.

Chapter 8 presents learning outcomes based on the researcher's experience utilising the user-centred design approach adopted to develop the mobile application.

Finally, Chapter 9 concludes the dissertation by summarising main findings from the research and outlining future work identified as a consequent of completing the reported research programme.

Chapter 2: Literature Review

2.1 Introduction

According to the bullying prevention charity, Ditch The Label, an estimated 5.43 million young people in the UK have been exposed to online bullying in the last year, and more than a million young people are subjected to severe cyberbullying daily (Ditch The Label, 2020). A staggering 69% of reported cyberbullying incidents occurred on Online Social Networks (OSN) like Facebook, Twitter and Instagram (Ofcom Research, 2019) establishing Social Media Platforms (SMP) as the dominant means through which online abuse is perpetuated

Today, there are two key aspects to cyberbullying: sociological and technological. The sociological aspects of cyberbullying encompass areas such as the definition of cyberbullying, its prevalence amongst different segments of society, its impact and predictors, and the ways by which it can be mitigated and prevented. The technological aspects of cyberbullying arise due to the use of technology as its primary means of perpetuation and research in this area focuses on developing automated means to identify, prevent and mitigate its occurrence.

This chapter details a review of relevant literature on the critical issues concerning cyberbullying, its detection and prevention using automated means. Section 2.2 begins by examining the sociological aspects of cyberbullying including its definitions, prevalence and the effects of gender, race and age on its prevalence. Sections 2.3 and 2.4 then examine cyberbullying from a technological perspective, beginning with a review of the literature on the techniques used to automatically detect cyberbullying in Section 2.3 and the various applications, systems and tools designed for cyberbullying prevention in Section 2.4. Section 2.4 also reports on the results of a survey of mobile applications

(currently available on the mobile app stores ¹²) that can aid the prevention and mitigation of cyberbullying that was conducted as part of this research agenda. Finally, Section 2.5 provides a summary of the chapter.

2.2 Definition, Prevalence and Consequences of Cyberbullying

Cyberbullying is “*an aggressive, intentional act carried out by a group or individual using electronic forms of contact, repeatedly against a victim that cannot easily defend him or herself*” (Smith *et al.*, 2008, pg. 376). It is wilful and repeated harm inflicted through the use of electronic devices (Hinduja and Patchin, 2008) and is often characterised by the posting of comments online to defame an individual, including the public disclosure of private information to inflict intentional emotional distress on the victim (Willard, 2005). Berger (2007) (as cited in Abeele and Cock 2013, pg. 95) distinguishes two types of cyberbullying: direct and indirect cyberbullying. He identified direct cyberbullying as when a victim is engaged directly by the bully (typified by sending explicit offensive and aggressive content) and indirect cyberbullying as a relational form of cyberbullying that involves subtler forms of abuse such as social exclusion and the use of sarcasm to ridicule.

While cyberbullying is regarded as a distinct and separate entity from traditional bullying (Guerin and Hennessy, 2002; Wingate *et al.*, 2013; Hinduja and Patchin, 2012; Law *et al.*, 2012), Olweus (2012) challenged this notion, claiming instead that cyberbullying should be classified as a subset of traditional bullying in the same manner as physical (bullying via physical violence or the threat of it) and relational (bullying someone by manipulating the social networks around them) bullying. He argued that online abuse could only be considered bullying if it satisfies three key conditions, namely: (i) the presence of a deliberate intent to harm; (ii) repeated acts of aggression; and (iii) an imbalance of power between bully and victim (Solberg and Olweus, 2003).

This argument, however, ignores the crucial role played by technology in the perpetuation of cyberbullying. Technology introduces additional elements to cyberbullying, many of

¹apple.com/uk/app-store

²play.google.com

which are irrelevant to traditional bullying and, as such, the three conditions above cannot always be applied in the same strict manner. For example, uploading an embarrassing picture or video to ridicule a person could qualify as cyberbullying due to its availability for repeated viewings (and causing ongoing distress to the victim) even though the originating act is a single isolated incident. Similarly, ascertaining a power differential within an online interaction can be difficult as the traditional roles of oppressor and the oppressed are more fluid in online social networks as observed in the many reports of high profile individuals being subjected to online abuse (BBC News, 2020). The cyberbullying definition by Hinduja and Patchin (2008) as **wilful and repeated harm perpetrated via electronic means** is, therefore, more reflective of technology's influence on cyberbullying and, as such, is the definition adopted by this thesis. Crucially, this definition identifies the deliberate and repeated nature of cyberbullying, which are fundamental characteristics of the phenomenon while relaxing the power differential requirement, an element of bullying that is quite difficult to measure in electronic communications.

An implication of the differences in cyberbullying definitions can be seen in the cyberbullying prevalence rate reported by many studies. The cyberbullying definition adopted by a given study and how cyberbullying questions are framed in surveys could impact reported frequencies. For example, in Coelho *et al.* (2016), respondents were provided with a small list of specific cyberbullying examples, and the study reported a lower victimisation rate (10%) compared to that of Zhou *et al.* (2013) which provided respondents with an extensive list of cyberbullying examples and discovered a higher victimisation rate of 57%. Equally, the classification of a cyberbullying victim by Olweus (2012) as someone that has been bullied electronically at least "once in the past four weeks" resulted in a lower victimisation rate (4%) compared to the 62% victimisation rate reported by Gkiomisi *et al.* (2017) that defined a victim as someone that has experienced at least one incident of online abuse in the previous six months. Furthermore, the time reference used with these questions (e.g., the six months time reference utilised by Gkiomisi *et al.* (2017) compared to three months used by Olweus (2012)) plays a significant role in the number of cyberbullying incidents reported.

Interestingly, a similar trend was not observed with regards to cyberbullying offending rates. While studies such as those by Lianos and McGrath (2018) and Zhou *et al.* (2013) which provided survey respondents with several typical examples of abusive behaviours

reported high offending rates of 80% and 35%, respectively, other studies such as those by Machimbarrena and Garaigordobil (2018) and Twardowska-Staszek *et al.* (2018) which used similar survey instruments reported lower perpetration rates (0.7% and 5.2%, respectively). A possible explanation for the low rates reported by Machimbarrena and Garaigordobil (2018) and Twardowska-Staszek *et al.* (2018) could be a tendency for respondents to provide desirable answers when questioned on behaviours they believe are socially unacceptable (Hinduja and Patchin, 2013) and/or could reflect the sample's demographic distribution (Slonje and Smith, 2008).

While age, gender, race, and sexual orientation are believed to impact reported frequencies of cyberbullying (Griezel *et al.*, 2012; Young and Govender, 2018; Hong *et al.*, 2016; DeSmet *et al.*, 2018b), substantial relationships between these demographic parameters and cyberbullying are yet to be consistently established. As girls demonstrated a higher involvement in relational forms of bullying such as social exclusion and emotional harassment (Hinduja and Patchin, 2008), the expectation is that they may similarly exhibit a higher cyberbullying perpetration rate due to the similarity in the ways both cyberbullying and relational bullying are perpetrated. The relationship between gender and cyberbullying is quite remarkable in its inconsistency. While studies such as those by Baldry *et al.* (2016), Gao *et al.* (2016), Guo2016, Li (2007), Beckman *et al.* (2013), Eden *et al.* (2013), Fanti *et al.* (2012) and Calvete *et al.* (2010) reported that girls are more likely to be victims of cyberbullying compared to boys, others like those of Snell and Englander (2010), Kowalski and Limber (2007), and Navarro and Jasinski (2013) discovered increased involvement of girls as both victims and perpetrators and some – Williams and Guerra (2007); Griezel *et al.* (2012); Hinduja and Patchin (2008); Tokunaga (2010) – found no gender differences at all in cyberbullying. The above studies, however, do not account for previous experience as victims of traditional bullying, and when this was taken into account, boys were found more likely to become cyberbullies than girls (Ágnes Zsila *et al.*, 2019).

The age of participants and the type of online social networks (OSNs) studied could also introduce variations in findings. For example, boys were found to experience more bullying on video game chatrooms while girls reported being bullied more on social media sites like Facebook and Instagram (Foody *et al.*, 2019). This suggests that gender proclivity for certain types of online social activities and the OSNs on which cyberbullying is being

studied may influence the gender distribution observed in cyberbullying victimisation and perpetration.

Race and ethnicity are common themes used to perpetuate cyberbullying (Dinakar *et al.*, 2011) and yet only a handful of studies examined the connection between race, ethnicity and cyberbullying. Consequently, much is still unknown about the nature of this relationship. While studies such as those of Lee and Shin (2017) and Edwards *et al.* (2016) discovered low prevalence rates for both cyberbullying victimisation and perpetration amongst Asian, Hispanic and African Americans, others like those of Hong *et al.* (2016) found African Americans are more likely to be bullied online compared to European Americans. Goebert *et al.* (2011) also reported a higher victimisation rate amongst ethnicities like Filipinos and Caucasians compared to Samoans and native Hawaiians in a Hawaiian high school sample. Similarly, Yousef and Bellamy (2015) reported that Arab Americans experienced cyberbullying victimisation more than African, Hispanic or European Americans in American middle and high schools. This implies that children from ethnicities considered foreign to the native culture are at a higher risk of cyberbullying victimisation.

Much as for race and ethnicity, research investigating the relationship between sexual orientation and cyberbullying is limited. DeSmet *et al.* (2018b) found higher levels of both traditional bullying victimisation and cyberbullying offending amongst lesbian, bisexual, gay and transgender (LGBT) youths compared to non-LGBT youths. Conversely, other studies such as those by Ybarra *et al.* (2015) and Elipe *et al.* (2018) found no such or any substantial relationship.

The prevailing cyberbullying victimisation and offending rates as reported by several studies (Tokunaga, 2010; Young and Govender, 2018; Kowalski and Limber, 2007; Dehue *et al.*, 2008; Slonje and Smith, 2008; Ybarra and Mitchell, 2008) suggest that young people's involvement with cyberbullying increases with age up to ages 15 and 16 years, after which a sharp decline is observed. This initial increase could be fuelled by improved access to electronic devices (Slonje and Smith, 2008) and reduced parental supervision as children grow older. The reduction in cyberbullying-related behaviours could then be due to changing and expanding interests as young people transition out of teenagehood and prepare for life after secondary school education.

With regards its effects on young people, cyberbullying has been consistently linked to many adverse outcomes including depression, anxiety, loneliness, low self-esteem, suicide ideation, substance abuse and poor academic performance (Dehue, 2013; Hinduja and Patchin, 2010; Berne *et al.*, 2014; Hoff and Mitchell, 2009; Şahin, 2012; Fahy *et al.*, 2016; Kim *et al.*, 2017; Wright, 2016; Khine *et al.*, 2020; Martínez-Monteagudo *et al.*, 2020). It breeds a desire for revenge in male adolescents (Caetano *et al.*, 2016), causes social media exhaustion and abandonment in young people (Cao *et al.*, 2019), and negatively impacts young people's overall engagement with ICT (Camacho *et al.*, 2018). Hinduja and Patchin (2019) found that adolescents bullied in school or online were significantly more likely to have suicidal thoughts, and those that have experienced both forms of bullying were more likely to attempt suicide. As cyberbullying victims are also often victims of traditional bullying, the extent to which these adverse effects can be attributed to cyberbullying has been questioned (Olweus and Limber, 2018). Kim *et al.* (2018), however, established that, after controlling for traditional bullying, cyberbullying was still found to be a significant predictor for emotional and behavioural problems. Smith (2015) posited that only via longitudinal studies such as that of Gámez-Guadix *et al.* (2013) can the isolated effects of cyberbullying be determined. Gámez-Guadix *et al.* (2013) conducted a 6-month longitudinal study and found that participants who experienced cyberbullying at the start of the study were more likely to suffer from depression by the end of the six months compared to those that were not bullied online. Likewise, Pabian and Vandebosch (2016) discovered that cyberbullying victims in their sample became perpetrators within six months of being abused online as part of an elaborate cycle of victimisation and offending.

A lack of affective empathy (Renati *et al.*, 2012; Çiğdem Topcu and Özgür Erdur-Baker, 2012), moral disengagement (Yang *et al.*, 2018), and problematic Internet use (e.g., visiting inappropriate websites) (Gámez-Guadix *et al.*, 2016) have been positively correlated with cyberbully offending. Cyberbullying has also been found to be more prevalent in children with disruptive home lives (Subrahmanyam and Greenfield, 2008) and is strongly influenced by the behaviours of peers and authoritative adults in children's lives (Hinduja and Patchin, 2013). Poor parental communication (Wang *et al.*, 2009) and restrictive supervision (Sasson and Mesch, 2014) have been found to increase the risk of cyberbullying involvement while children that enjoy constructive relationships with parents and guardians were found to be less likely to be involved in cyberbullying as victims or

perpetrator (Doty *et al.*, 2018).

Despite receiving considerable research attention in recent years, there is still much that is unknown about cyberbullying. Compared to traditional bullying, it is still a relatively new area of study (as can be attested to by the many inconsistent findings). As its perpetration is facilitated by technology, it is not inconceivable that the lasting solutions to its mitigation and prevention may also very well rely on technology. Indeed, the automated detection of cyberbullying and online abuse has become a key subfield of Natural Language Processing (NLP) and one that has witnessed considerable growth, as is highlighted in the following sections.

2.3 Cyberbullying Detection Techniques

Schools' responses to cyberbullying are typically aimed at raising awareness, fostering positive values and interpersonal relationships amongst young people, as well as the creation of policies and guidelines to govern the use of electronic devices for extracurricular activities (Ortega-Ruiz *et al.*, 2012; Cassidy *et al.*, 2018; Doty *et al.*, 2018). Under the Education and Inspection Act 2006³, schools in the United Kingdom are required to provide a safe and healthy environment for all students, and this includes a documented anti-bullying policy. While cyberbullying is covered within this, the reality is that many schools are not well equipped to deal with cyberbullying and quite often there is a disconnect between what teachers and school administrators know about cyberbullying and its reality amongst students. Schools' assemblies, technology use guidelines, smart device usage restrictions and anti-cyberbullying posters are some of the tools used by educators to mitigate cyberbullying.

As noted, it is typically the case that teachers have not been suitably trained and educated on cyberbullying, its mitigation and prevention (BBC News, 2019). Consequently, it often remains undealt with by schools and exists as an underbelly to ordinary school experiences. The pervasive nature of cyberbullying also means that it frequently occurs outside school boundaries, rendering educators somewhat powerless in its mitigation.

³legislation.gov.uk/ukpga/2006/40/contents

While initial technological efforts to combat cyberbullying have also sought to emulate the same fundamental strategies used by schools (typically in the form of electronic educational material or videos), there is now a paradigmatic shift to actively tackle cyberbullying by applying computing techniques such as Machine Learning (ML) to detect and prevent it. In this regard, the detection of cyberbullying and online abuse can be formulated as a text classification problem that seeks to separate abusive and offensive content from a larger text corpus.

Cyberbullying detection can, therefore, be formally defined as the automated identification of bullying attacks perpetrated via electronic media, and it is concerned with performing one or more of the following tasks:

- i. identification of online harassment;
- ii. computing the severity of the bullying incident;
- iii. identification of the roles involved in a cyberbullying incident; and
- iv. the classification of resulting events that occur after a cyberbullying incident (e.g., detecting the emotional state of a victim after receiving a bullying message).

Identifying online harassment involves detecting offensive language in messages and posts and is sometimes required as an initial stage before further cyberbullying detection tasks can be performed. The presence of offensive content within a message can be used as the basis of computing a score symptomatic of the bullying severity, and it is often the next in a sequence of cyberbullying detection activities following the successful identification of cyberbullying attacks. In (iii) – roles identification – the aim is to decipher how the actions of the parties involved affect the cyberbullying incident outcome and to use this as the basis to assign bullying roles. Elements of the three tasks above are then required in (iv) to associate the cyberbullying attack to seemingly unrelated events (e.g., associating past cyberbullying incidents to the manifestation of suicide ideation behaviour on social media).

An extensive literature search was conducted across Scopus, ACM, IEEE Xplore and Google Scholar digital libraries to discover relevant academic literature on the automated detection of cyberbullying, online anti-social behaviour and harassment. In total, 105 papers were examined in detail as part of the literature review (Table 2.1)

The literature on cyberbullying and offensive language detection can be broadly categorised into two groups: those that use traditional machine learning approaches and those that utilise deep-learning methods. Traditional machine learning approaches make use of techniques such as supervised and semi-supervised learning; they represented the state-of-the-art in offensive language detection for many years but have now been largely surpassed by deep-learning methods. The following subsections discuss both traditional ML and deep learning approaches to cyberbullying detection. Transformer-based methods, a sub-category of deep learning approaches that has recently emerged as the new state-of-the-art in many NLP tasks, are also discussed in a separate subsection.

Study	Tasks Performed				Approach					Data Source									Classifiers	Language	
	Detecting offensive content	Detecting bullying roles	Detecting bullying severity	Detecting bullying aftermath	Supervised Learning	Lexicon-Based	Rules-Based	Mixed Initiatives	Deep Learning	Twitter	Facebook	Instagram	YouTube	MySpace	Sping.me	Ask.fm	Wikipedia	Vine			Others/Unknown
Mahmud <i>et al.</i> , 2008	✓						✓												✓	-	English
Yin <i>et al.</i> , 2009	✓				✓														✓	SVM	English
Bosse and Stam, 2011		✓																	✓	-	English
Dinakar <i>et al.</i> , 2011	✓				✓							✓								Naïve Bayes, SVM, J48, JRip	English
Sanchez and Kumar, 2011	✓	✓			✓					✓										Naïve Bayes	English
Serra and Venter, 2011	✓						✓												✓	Neural Networks	English
Burn-Thorton and Burman, 2012	✓				✓														✓	kNN	English
Chen <i>et al.</i> , 2012		✓	✓				✓					✓								Naïve Bayes, SVM	English
Dadvar and De Jong, 2012	✓				✓								✓							SVM	English
Dadvar <i>et al.</i> , 2012a	✓				✓								✓							SVM	English
Dadvar <i>et al.</i> , 2012b	✓			✓	✓								✓							SVM	English
Dinakar <i>et al.</i> , 2012	✓						✓					✓		✓						Naïve Bayes, SVM, J48, JRip	English

Study	Tasks Performed				Approach					Data Source									Classifiers	Language	
	Detecting offensive content	Detecting bullying roles	Detecting bullying severity	Detecting bullying aftermath	Supervised Learning	Lexicon-Based	Rules-Based	Mixed Initiatives	Deep Learning	Twitter	Facebook	Instagram	YouTube	MySpace	Sping.me	Ask.fm	Wikipedia	Vine			Others/Unknown
Nahar <i>et al.</i> , 2012	✓	✓			✓														✓	SVM	English
Perez <i>et al.</i> , 2012			✓			✓													✓	-	English
Sood and Churchill, 2012a	✓				✓														✓	SVM	English
Sood and Churchill, 2012b	✓				✓														✓	SVM	English
Xu <i>et al.</i> , 2012a	✓	✓			✓					✓										Naïve Bayes, SVM, Logistic Regression, LDA, Conditional Random Fields (CRF)	English
Xu <i>et al.</i> , 2012b	✓				✓					✓							✓			Naïve Bayes, SVM, Logistic Regression, LDA	English
Dadvar <i>et al.</i> , 2013a			✓					✓				✓								MCES	English
Dadvar <i>et al.</i> , 2013b			✓		✓							✓								SVM	English
Kontostathis, 2013	✓				✓									✓						EDLSI	English
Munezero, 2013	✓				✓													✓		Naïve Bayes,	English

Study	Tasks Performed				Approach					Data Source									Classifiers	Language		
	Detecting offensive content	Detecting bullying roles	Detecting bullying severity	Detecting bullying aftermath	Supervised Learning	Lexicon-Based	Rules-Based	Mixed Initiatives	Deep Learning	Twitter	Facebook	Instagram	YouTube	MySpace	Sping.me	Ask.fm	Wikipedia	Vine			Others/Unknown	
																				SVM, J48		
Nahar <i>et al.</i> , 2013	✓	✓			✓															✓	SVM, LDA, HITS	English
Sheeba and Vivekanandan, 2013	✓				✓					✓	✓									✓	Maximum Entropy, Fuzzy Systems	English
Bretschneider <i>et al.</i> , 2014	✓						✓			✓											-	English
Dadvar <i>et al.</i> , 2014		✓						✓					✓								Naïve Bayes, SVM, C.45, MCEs	English
Del Bosque and Garza, 2014			✓		✓					✓											Multi-Layer Perceptron (MLP) Neural Network	English
Fahrnberger <i>et al.</i> , 2014	✓					✓														✓	-	English
Huang <i>et al.</i> , 2014	✓				✓									✓						✓	Radom Forrest, K-FSVM, Naïve Bayes, Logistic Regression	English
Munezero, 2014	✓				✓									✓							SVM (Linear)	English
Nahar <i>et al.</i> , 2014	✓				✓															✓	Naïve Bayes, SVM, J48	English

Study	Tasks Performed				Approach					Data Source									Classifiers	Language	
	Detecting offensive content	Detecting bullying roles	Detecting bullying severity	Detecting bullying aftermath	Supervised Learning	Lexicon-Based	Rules-Based	Mixed Initiatives	Deep Learning	Twitter	Facebook	Instagram	YouTube	MySpace	Sping.me	Ask.fm	Wikipedia	Vine			Others/Unknown
Parime and Suri, 2014	✓				✓														✓	SVM, Multi-Layer Perceptron (MLP) Neural Network	English
Potha and Maragoudakis, 2014	✓		✓		✓														✓	SVM, Logistic Regression	English
Chavan and Shylaja, 2015	✓				✓					✓										SVM-PolyKernel, J48, SVM-NormalizedPolyKernel, RandomForest	English
Galán-García <i>et al.</i> , 2015		✓			✓					✓										J48, Naïve Bayes, SMO, ZeroR	English
Hosseinmardi <i>et al.</i> , 2015	✓				✓						✓					✓				Naïve Bayes, SVM	English
Mangaonkar <i>et al.</i> , 2015	✓				✓					✓										Naive Bayes (NB), Logistic regression, SVM	English
NaliniPriya and Asswini, 2015	✓				✓														✓		English
Nandhini and Sheeba, 2015a	✓				✓									✓	✓					Naïve Bayes	English

Study	Tasks Performed				Approach					Data Source								Classifiers	Language		
	Detecting offensive content	Detecting bullying roles	Detecting bullying severity	Detecting bullying aftermath	Supervised Learning	Lexicon-Based	Rules-Based	Mixed Initiatives	Deep Learning	Twitter	Facebook	Instagram	YouTube	MySpace	Sping.me	Ask.fm	Wikipedia			Vine	Others/Unknown
Nandhini and Sheeba, 2015b	✓				✓									✓	✓					Naïve Bayes	English
Rafiq <i>et al.</i> , 2015	✓				✓													✓		Naïve Bayes, AdaBoost, Decision-Tree, Random Forest	English
Rajadesingan <i>et al.</i> , 2015	✓						✓			✓										-	English
Squicciarini <i>et al.</i> , 2015		✓		✓	✓									✓	✓					C4.5 Decision Tree	English
Al-Garadi <i>et al.</i> , 2016	✓				✓					✓										Naive Bayes, SVM, Random Forest, KNN	English
Hosseinmardi <i>et al.</i> , 2016	✓				✓						✓									Logistic Regression	English
Potha <i>et al.</i> , 2016	✓				✓													✓		SVM	English
Rafiq <i>et al.</i> , 2016	✓				✓													✓		AdaBoost, Decision Tree, Random Forest, SVM Linear, SVM Polynomial, SVM RBF, SVM Sigmoid,	English

Study	Tasks Performed				Approach					Data Source									Classifiers	Language	
	Detecting offensive content	Detecting bullying roles	Detecting bullying severity	Detecting bullying aftermath	Supervised Learning	Lexicon-Based	Rules-Based	Mixed Initiatives	Deep Learning	Twitter	Facebook	Instagram	YouTube	MySpace	Sping.me	Ask.fm	Wikipedia	Vine			Others/Unknown
																				KNN, Naive Bayes, Perceptron, Ridge classifier, Logistic Regression.	
Singh <i>et al.</i> , 2016	✓				✓					✓										Probabilistic Information Fusion	English
Waseem and Hovy, 2016	✓				✓					✓										Logistic Regression	English
Zhang <i>et al.</i> , 2016	✓						✓			✓				✓						PCNN	English
Zhong <i>et al.</i> , 2016	✓				✓		✓				✓									SVM, Random Forest, CNN	English
Dani <i>et al.</i> , 2017	✓				✓					✓			✓							Custom	English
Davidson <i>et al.</i> , 2017	✓				✓					✓										Logistic Regression, Linear SVM	English
Foong and Oussalah, 2017	✓				✓										✓					SVM	English
Ptaszynski <i>et al.</i> , 2017	✓						✓											✓		CNN	Japanese
Sarna and Bhatia, 2017	✓				✓					✓										SVM, K-Nearest Neighbour, Naive Bayes, Decision	English

Study	Tasks Performed				Approach					Data Source									Classifiers	Language		
	Detecting offensive content	Detecting bullying roles	Detecting bullying severity	Detecting bullying aftermath	Supervised Learning	Lexicon-Based	Rules-Based	Mixed Initiatives	Deep Learning	Twitter	Facebook	Instagram	YouTube	MySpace	Sping.me	Ask.fm	Wikipedia	Vine			Others/Unknown	
																				Trees.		
Ting <i>et al.</i> , 2017	✓				✓					✓												English
Zhao and Mao, 2017	✓				✓					✓				✓							SVM (Linear)	English
Zhao <i>et al.</i> , 2017	✓				✓					✓											SVM (Linear)	English
Agrawal and Awekar, 2018	✓							✓		✓					✓		✓				CNN, LSTM, BLSTM, BLSTM with Attention	English
Bu and Cho 2018	✓							✓										✓			CNN + LRCN	English
Chandra <i>et al.</i> , 2018	✓							✓		✓											RNN	English
Chen <i>et al.</i> , 2018	✓				✓					✓	✓								✓		SVM Linear, K-Nearest Neighbour, Random Forest, Logistic Regression Stochastic Gradient Descent, Naïve Bayes	English
Chen <i>et al.</i> , 2018	✓							✓		✓											CNN	English
Gorro <i>et al.</i> , 2018	✓				✓						✓										SVM	English

Study	Tasks Performed				Approach					Data Source									Classifiers	Language	
	Detecting offensive content	Detecting bullying roles	Detecting bullying severity	Detecting bullying aftermath	Supervised Learning	Lexicon-Based	Rules-Based	Mixed Initiatives	Deep Learning	Twitter	Facebook	Instagram	YouTube	MySpace	Sping.me	Ask.fm	Wikipedia	Vine			Others/Unknown
Ibrahim <i>et al.</i> , 2018	✓							✓									✓			LSTM+CNN+GRU	English
Lee <i>et al.</i> , 2018	✓					✓				✓										-	English
Pawar <i>et al.</i> , 2018	✓				✓									✓						Multinomial Naïve Bayes, Stochastic Gradient Descent	English
Ptaszynski <i>et al.</i> , 2018	✓				✓													✓		Custom algorithm	English
Raisi and Huang, 2018a		✓				✓				✓		✓				✓				-	English
Raisi and Huang, 2018b	✓	✓			✓			✓		✓		✓				✓				LSTM, BOW, word embeddings and doc2vec, node2vec	English
Rakib and Soon, 2018	✓				✓													✓		Random Forrest	English
Rosa <i>et al.</i> , 2018	✓							✓						✓						CNN	English
Van Hee <i>et al.</i> , 2018	✓				✓											✓				SVM Linear	English, Dutch
Ventirozos <i>et al.</i> , 2018	✓				✓								✓							Hidden Markov Model, SVM	English

Study	Tasks Performed				Approach					Data Source									Classifiers	Language	
	Detecting offensive content	Detecting bullying roles	Detecting bullying severity	Detecting bullying aftermath	Supervised Learning	Lexicon-Based	Rules-Based	Mixed Initiatives	Deep Learning	Twitter	Facebook	Instagram	YouTube	MySpace	Sping.me	Ask.fm	Wikipedia	Vine			Others/Unknown
Yao <i>et al.</i> , 2018; 2019	✓				✓							✓								Custom CONcISE	English
Zois <i>et al.</i> , 2018	✓				✓					✓										Custom	English
Aci <i>et al.</i> , 2019	✓				✓							✓	✓	✓						MLP, Logistic Regression, Stochastic Gradient Descent SGD	English
Anand and Eswari 2019	✓						✓											✓		LSTM, CNN	English
Andleeb <i>et al.</i> , 2019	✓				✓								✓						✓	SVM, Naïve Bayes	English
Anindyati <i>et al.</i> , 2019	✓							✓	✓											CNN, LSTM, Bi-LSTM	Indonesia
Anitha <i>et al.</i> , 2019	✓				✓				✓											Naïve Bayes	English
Balakrishnan <i>et al.</i> , 2019		✓			✓				✓											Random Forrest	English
Banerjee <i>et al.</i> , 2019	✓						✓		✓											CNN	English
Chatzakou <i>et al.</i> , 2019		✓			✓		✓		✓											Naïve Bayes, LADTree, LMT, NBTree, Random Forest, Functional Tree; Random	English

Study	Tasks Performed				Approach					Data Source									Classifiers	Language	
	Detecting offensive content	Detecting bullying roles	Detecting bullying severity	Detecting bullying aftermath	Supervised Learning	Lexicon-Based	Rules-Based	Mixed Initiatives	Deep Learning	Twitter	Facebook	Instagram	YouTube	MySpace	Sping.me	Ask.fm	Wikipedia	Vine			Others/Unknown
																				Forest + BayesNet + Naive Bayes + AdaBoost, RNN	
Cheng <i>et al.</i> , 2019		✓			✓					✓										Custom	English
Cheng <i>et al.</i> , 2019	✓						✓				✓									GRU-RNN	English
Cheng <i>et al.</i> , 2019b		✓			✓						✓							✓		Logistic Regression, Random Forest SVM Linear	English
Gutiérrez- Esparza <i>et al.</i> , 2019	✓				✓						✓									OneR, Radom Forest, Variable Importance Measures VIM	Spanish
Haidar <i>et al.</i> , 2019	✓				✓					✓										SVM, K- Nearest Neighbour, Random Forest, Bayesian Logistic Regression, Stochastic Gradient Descent	Arabic

Study	Tasks Performed				Approach					Data Source									Classifiers	Language	
	Detecting offensive content	Detecting bullying roles	Detecting bullying severity	Detecting bullying aftermath	Supervised Learning	Lexicon-Based	Rules-Based	Mixed Initiatives	Deep Learning	Twitter	Facebook	Instagram	YouTube	MySpace	Sping.me	Ask.fm	Wikipedia	Vine			Others/Unknown
Kumari <i>et al.</i> , 2019	✓							✓											✓	CNN	English
Li, 2019	✓					✓													✓	-	Chinese
Mouheb <i>et al.</i> , 2019	✓				✓					✓			✓							Naïve Bayes	Arabic
Novalita <i>et al.</i> , 2019	✓				✓					✓										Random Forest	English
Ousidhoum <i>et al.</i> , 2019	✓				✓				✓	✓										Logistic Regression, Bi-LSTM	English, French, Arabic
Saha and Senapati, 2019	✓							✓		✓	✓									LSTM	English, German, Hindi
Singh and Kaur 2019	✓				✓					✓					✓	✓				Cuckoo Search + SVM	English
Tomkins <i>et al.</i> , 2019	✓	✓			✓					✓										SVM	English
Zhong <i>et al.</i> , 2019	✓							✓			✓									CNN	English
Chatterjee and Das, 2020				✓	✓					✓										-	English
Kargutkar and Chitre, 2020	✓								✓	✓										CNN	English
Niu <i>et al.</i> , 2020	✓								✓										✓	Bi-LSTM	Chinese
Purnamasari <i>et al.</i> , 2020	✓				✓					✓										SVM	English

Study	Tasks Performed				Approach					Data Source									Classifiers	Language	
	Detecting offensive content	Detecting bullying roles	Detecting bullying severity	Detecting bullying aftermath	Supervised Learning	Lexicon-Based	Rules-Based	Mixed Initiatives	Deep Learning	Twitter	Facebook	Instagram	YouTube	MySpace	Sping.me	Ask.fm	Wikipedia	Vine			Others/Unknown
Van Bruwaene <i>et al.</i> , 2020	✓				✓				✓	✓	✓	✓							✓	SVM, CNN, XGBoost	English
Wu <i>et al.</i> , 2020	✓				✓														✓	fastText	Chinese

TABLE 2.1: Discovered literature on cyberbullying detection and the approaches used.

2.3.1 Traditional Machine Learning Approaches

Yin *et al.* (2009) appear to have pioneered work utilising supervised learning techniques for cyberbullying detection. They theorised that, due to the low number of harassment posts within a text corpus, a harassment post would appear significantly different from its neighbouring posts. On this basis, they used a document's immediate neighbourhood of k posts ($k = 3$) as a classification feature to identify cyberbullying documents as those exhibiting substantial difference from their neighbours. They used a Support Vector Machine (SVM) classifier to analyse posts from social websites like Slashdot⁴, Kongregate⁵, and MySpace⁶. A Support Vector Machine is a discriminative algorithm that uses provided training data to output a hyperplane that separates the data into different classes (Suthaharan, 2016). As offensive documents within a corpus share similar characteristics, they will naturally be segregated by the hyperplane, effectively classifying them as members of an *offensive* class. Since then, supervised learning has emerged as the most frequently used machine learning technique for cyberbullying detection (Salawu *et al.*, 2017). The aim in supervised learning is the inference of a mapping function from a labelled set of data (referred to as training data) such that

$$Y = f(X)$$

where X and Y are input and output variables, respectively. The mapping function makes use of learned characteristics of the data (called features) to predict Y for new instances of X . A feature is a measurable property of the data, for example, age, employment status, credit score, and income level could be characteristics of a dataset used by a model to make a mortgage decision. An integral part of training traditional ML models is identifying the features to represent the data. These features often require tuning and enhancing to derive the most benefits from them; this process is typically referred to as feature engineering and can be as important to supervised learning as the choice of algorithm used.

Unsupervised learning utilises algorithms that model the underlying structure of data to learn more about it without a training sample, and semi-supervised learning techniques

⁴Slashdot.org

⁵Congregate.com

⁶myspace.com

use elements of both supervised and unsupervised learning. Semi-supervised learning is used when only a small subset of a large amount of data is labelled and involves first using unsupervised learning techniques to discover and learn the structure of the data. Supervised learning methods are then used to make best guess predictions for a sample of the unlabelled data and the predictions fed back into the learning algorithm as training data to develop a model for predicting the rest of the unseen data.

Following on from the work of Yin *et al.* (2009), other researchers like Dinakar *et al.* (2011); Dadvar and Jong (2012); Dadvar *et al.* (2012a,b); Choudhury *et al.* (2013); Sood *et al.* (2012a); Munezero *et al.* (2013); Nahar *et al.* (2014); Hosseinmardi *et al.* (2015); Zhao *et al.* (2016) and Foong and Oussalah (2017) expanded on the use of SVMs for detecting various forms of harassment and cyberbullying. Aside from SVMs, other machine learning algorithms used for cyberbullying detection include Naïve Bayes, Logistic Regression, Decision Tree and Random Forests. Naïve Bayes is a probabilistic classifier that classifies data by finding models that segregate data based on the assumption that the presence of one feature in a class is unrelated to the presence of any other features (Rish, 2001). Logistic regression is a statistical method for analysing data that contain independent variables that result in one of two possible outcomes (Dreiseitl and Ohno-Machado, 2002). For example, in a mortgage application, an applicant's age can make them ineligible if they are younger than 18 years irrespective of their income. A Decision Trees algorithm repeatedly splits data into a tree-like structure (hence the name) according to a function that continuously maximises the separation of the data in such a way that different classes of data are in different branches (Breiman *et al.*, 2017). Random Forests is an ensemble learning method that creates multiple decision trees during training and predicts an output by combining the results of individual trees (Ho, 1995). These algorithms enjoyed huge popularity in cyberbullying detection research and, as previously mentioned, represented the state-of-the-art in cyberbullying detection for many years but are now more likely to be used as baseline classifiers.

Dinakar *et al.* (2011) improved on the work of Yin *et al.* (2009) by segregating the training data into smaller clusters based on themes such as racism, culture, sexuality, and intelligence. They theorised that the performance of an SVM classifier could be improved by improving the homogeneity of the training data, resulting in improved performance by the classifiers on the individual clusters compared to the superset of all training data. The merits of this approach were corroborated by other researchers such as Dadvar and Jong

(2012); Dadvar *et al.* (2012a); Nahar *et al.* (2014) and Romsaiyud *et al.* (2017) who similarly segregated the training data before using it to train the classifier. The improvement achieved by this method is due to the specialism introduced into the trained model via the segregate training data, but this is achieved at the cost of the model's ability to generalise well on other topics. This exposed a critical failing of early work developing machine learning models for cyberbullying detection: the models developed did not generalise well when predicting unseen data from different domains. In attempting to resolve this common issue, researchers devised innovative techniques to improve the performance of classifiers. One such method is the use of feature selection algorithms like Chi2 (Liu and Setiono, 1995), ReliefF (Sikonja and Kononenko, 2003), MRMR (Peng *et al.*, 2005), and SVM-RFE (Huang *et al.*, 2014) to discover the most impactful features within the training data as seen in the work of Zois *et al.* (2018); Yao *et al.* (2018, 2019) and Çiğdem Aci *et al.* (2019). Trained models use these features to generate the mapping functions employed to categorise text; therefore, by focusing on the most critical features, a model can potentially improve its performance.

The use of voting functions (or ensemble learning) to combine the output of multiple models is another performance-enhancing technique used across many studies. As machine learning algorithms often excel at different aspects of the same task, combining the outputs in this way could potentially improve overall performance. It is a technique that is well used in many areas of NLP and has also gained traction in cyberbullying detection. The work by Sood *et al.* (2012a,b); Mangaonkar *et al.* (2015); Nandhini and Sheeba (2015) and Chavan and Shylaja (2015) are examples of studies that trained multiple models for cyberbullying detection and combined the outputs of the different models to achieve a better performance compared to that of a single model. While the additional computational processes required to generate multiple predictions and train multiple models makes the approach computationally expensive, it remains one of the key strategies used in sensitive predictions such as medical diagnosis (Khuriwal and Mishra, 2018) and financial decision-making (Randhawa *et al.*, 2018). The subjectivity involved in identifying cyberbullying could also introduce sensitivity to its automated detection, and therefore, ensemble learning methods are similarly used to reduce the possibility of false positives.

As previously mentioned, machine learning models represent data as a function of the observed features, and the number of features used can vary from a few to tens of

thousands. Conventionally, ML models developed for cyberbullying detection have used feature extraction techniques like Bag of Words (BoW), TF-IDF (Term Frequency Inverse Document Frequency) and N-grams to facilitate learning. BoW represents a document as the count of its words while TF-IDF is the product of the term frequency (i.e., frequency of a word – a raw word count in its simplest form) and inverse document frequency (a measure of how important the word is) for a document. An N-gram is simply a sequence of N-words, for example, “patient dog” and “fattest bone” are bi-grams of “the patient dog eats the fattest bone”. More recently, word embeddings as popularised by vector space models like word2vec (Mikolov *et al.*, 2013), GloVe (Pennington *et al.*, 2014), SSWE (Tang *et al.*, 2014) and fastText (Joulin *et al.*, 2016) have emerged as standard features used in text classification tasks. A word embedding is a real-valued vector presentation of a word such that words that have the same meaning have a similar representation (Bahdanau *et al.*, 2017). A vector space model represents documents as vectors with the vectors’ dimensions representing a word’s or phrase’s occurrence within the document. Vectors are typically generated on large volumes of data and then used as features in classifiers for downstream tasks such as offensive language and cyberbullying detection. Zhao *et al.* (2016) are early pioneers in the use of these models for cyberbullying detection and, more recently, studies like those of Rakib and Soon (2018) and (Wu *et al.*, 2020) have used word embeddings for similar purposes. These studies achieved superior performance in cyberbullying detection tasks compared to conventional textual representation like BoW, N-grams and TF-IDF and, since embeddings are typically generated from large corpora of generic text, the ensuing models demonstrated a better understanding of written language and are thus able to generalise better on different types of data.

The use of bullying wordlists (human-curated lists of offensive terms) can be found in some of the early work in cyberbullying detection research including that of Pérez *et al.* (2012); Fahrnberger *et al.* (2014) and Kontostathis *et al.* (2013). It is, therefore, interesting to discover its use in more recent work like that of Raisi and Huang (2018b); Lee *et al.* (2018); Hang and Dahlan (2019) and Li (2019). That the use of wordlists has endured for so long in cyberbullying detection research attests to its practicality in detecting online abuse. Its efficacy somewhat lies in its simplicity and, while the presence of an offensive term does not on its own indicate cyberbullying, when combined with other methods it can augment the models’ overall effectiveness.

The research discussed so far has predominantly focused on the use of what can be described as content-based features in the detection of cyberbullying. These are features generated from the actual content of messages. Cyberbullying detection is, however, not limited to the use of these types of features alone. Aside from the text content of messages, multi-modal information like image, video, user profile, number of followers, frequency of comments and uploads can be mined from social media data. Such information has been used successfully in studies such as those of Serra and Venter (2011); Nahar *et al.* (2014) and Squicciarini *et al.* (2015) as features for cyberbullying detection. It should, however, be noted that, while platform metadata like time and location of posting, number of posts, etc. are reliable data provided by the SMP, user profile information like age and gender is unreliable as it can be easily falsified and is not validated. The relative importance assigned to such user-provided information should, therefore, be carefully considered to avoid misrepresentation. This is likely why studies utilising multi-modal features (Dadvar *et al.*, 2013; Hosseinmardi *et al.*, 2015; Rafiq *et al.*, 2015; Al-Garadi *et al.*, 2016; Ting *et al.*, 2017; Cheng *et al.*, 2019) have favoured platform-generated data over user-provided ones.

As cyberbullying detection researchers continuously explored ways to improve the performance of traditional ML models, their work became more multi-disciplinary, and techniques were borrowed from other areas of NLP to advance the state-of-the-art in cyberbullying detection. Sentiment analysis, a sub-field of natural language processing that involves the analysis of opinions, sentiments, attitudes, and emotions expressed in written language (Liu, 2012), is one such discipline. It has been successfully used in areas such as detecting sentiments in informal product reviews on social media (Saif *et al.*, 2012) and analysing market trends in financial forecasting (Oliveira *et al.*, 2013) and has also found usage in cyberbullying detection.

In one of the notable works in this area (Munezero *et al.*, 2013), it was discovered that the use of sentiment-based features alone achieves lower performance than other feature types. This is perhaps not surprising as the presence of negative sentiments within a message is rarely a sufficient predictor for cyberbullying and harassment. These sentiments could have been expressed in jest or sarcastically, and this is a likely reason for why sentiment analysis techniques are often combined with other features like TF-IDF, BoW and N-grams (Sanchez and Kumar, 2011; Nahar *et al.*, 2013; Sheeba and Vivekanandan, 2013; v Bosque and Garza, 2014; Munezero *et al.*, 2014; Rafiq *et al.*, 2015; Squicciarini *et al.*, 2015) when used for cyberbullying detection. The improvement

in performance achieved in these studies by augmenting the detection process with such sentiment analysis techniques was marginal at best and perhaps not worth the efforts required to incorporate them. This could have contributed to the decline in the use of these techniques in cyberbullying detection.

As mention above, the sentiments expressed within a message should not always be interpreted literally as they could have been intended sarcastically. The ability to identify sarcasm is, therefore, a useful trait for a cyberbullying detection classifier to possess. Sarcasm detection is a sub-field of sentiment analysis that deals with the prediction of sarcasm in text and the work done in this area could prove very beneficial to cyberbullying detection. While sarcasm can be used to ridicule and by extension bully a person, it is interesting that none of the existing cyberbullying datasets include sarcasm and other less explicit forms of cyberbullying like social exclusion as labels. This highlights a critical area of cyberbullying detection research that is currently being overlooked and in need of more research attention. The techniques developed in studies such as those by Rajadesingan *et al.* (2015); Amir *et al.* (2016); Cai *et al.* (2019); Pant and Dadu (2020); Jena *et al.* (2020) and Potamias *et al.* (2020) to identify sarcasm in text have the potential to be explored and adapted to improve the abilities of cyberbullying classifiers to correctly detect when sarcasm is being used to perpetrate bullying.

Another area of cyberbullying detection that could benefit from more research is the development of novel mechanisms by which to identify a pattern of harassment within a conversation spanning several posts and messages. This involves identifying the temporal characteristics of messages and posts and using them to facilitate cyberbullying detection. Whilst studies by Potha and Maragoudakis (2015); NaliniPriya and Asswini (2015); Potha *et al.* (2016); Soni and Singh (2018) and Gupta *et al.* (2020) are notable works in this area, they simultaneously illustrate the paucity of research in this field. In their work detecting sexual predation and harassment, Potha and Maragoudakis (2015) and Potha *et al.* (2016) modelled offenders' messages as a time series while NaliniPriya and Asswini (2015) used ego networks to compute temporal changes in users' relationships, and used the detected changes as features for cyberbullying detection. Soni and Singh (2018) and Gupta *et al.* (2020) studied the temporal aspects of online abuse by examining properties like messaging intervals, frequency of posting, and the incident's duration.

Similar to the use of wordlists, expert and rule-based systems are two other non-ML methods explored by researchers to enhance the performance of cyberbullying detection classifiers. As these approaches allowed the injection of human-based knowledge into the detection process, researchers were able to study areas of cyberbullying detection that receive limited attention, such as the detection of subtler forms of cyberbullying. BullySpace – developed by Dinakar *et al.* (2012) – represents one of the earliest attempts at this; it uses a sparse matrix representation of assertions based on stereotypes derived from LGBT-related instances of bullying. The assertions are statements like “*lipstick is used by girls*”, “*lipstick is part of makeup*”, “*makeup is used by girls*” and so on. These assertions then allowed the system to infer that a statement such as “*did you go lipstick shopping with your mum today*” addressed to a heterosexual male can be classed as bullying. While it was certainly an innovative attempt at augmenting cyberbullying detection with real-world awareness, BullySpace’s reliance on a finite number of assertions limits its usefulness to a narrow range of scenarios since the available assertions cannot possibly cover the full range of bullying situations.

As cyberbullying detection researchers have become multi-disciplinary in their approach to innovate and improve performance, the implication is that traditional ML models are struggling to advance the state-of-the-art for cyberbullying detection. This has coincided with the increasing popularity of deep learning models in NLP and consequently has ushered in the present era of utilising deep-learning methods for cyberbullying detection.

2.3.2 Deep Learning Approaches

Deep learning is a computational technique for classifying patterns, based on sample data, using artificial neural networks with multiple layers (Marcus, 2018). Neural networks consist of a number of input units connected to multiple hidden layers (where the learning takes place) which are then connected to a set of output units. The term *deep* refers to the hidden layers of a neural network, thus the more hidden layers a network has the *deeper* it is (see FIGURE 2.1). Inputs to a neural network can be images, parts of an image, words or phrases. The network learns various relationships about its inputs and then uses this to make predictions when exposed to unseen data.

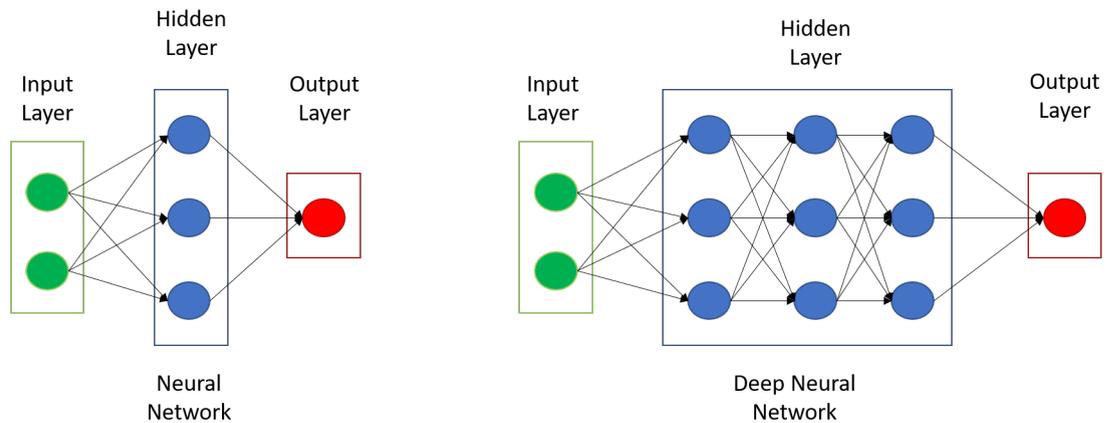


FIGURE 2.1: Neural Network vs Deep Neural Network.

While the historical roots of deep learning can be traced back to the late seventies and early eighties (Schmidhuber, 2015), it was not until 2012 when the use of Deep Neural Networks (DNN) achieved state-of-the-art results in the ImageNet object recognition competition (Krizhevsky *et al.*, 2017) that interest in its use was reignited along with recent computing discoveries such as the use of the computer's Graphical Processing Unit (GPU) – a component developed for games and 3D graphics – for mathematical computation.

A Convolutional Neural Network (CNN) is a deep neural network that uses convolution and pooling functions to analyse their inputs (Kalchbrenner *et al.*, 2014). The convolution function is an operation between a vector of weights and an input vector, essentially modifying the input vectors via a dot product. The pooling function is used to down-sample the input, reducing its dimensionality and allowing for assumptions to be made about properties of the discarded dimensions. CNN has been used in studies such as those by Ptaszynski *et al.* (2017); Agrawal and Awekar (2018); Rosa *et al.* (2018); Bu and Cho (2018); Anand and Eswari (2019); Zhong *et al.* (2019); Anindyati *et al.* (2019) and Chen and Li (2020) to detect online abuse, outperforming traditional ML classifiers such as SVM, Logistic Regression, Random Forest and Naïve Bayes in cyberbullying detection. While word embeddings have traditionally been used as inputs for CNN models (Banerjee *et al.*, 2019; Kargutkar and Chitre, 2020), studies such as those by Zhang *et al.* (2016); Chen *et al.* (2018) and Zhao *et al.* (2020) have identified limitations with their use and instead proposed alternative forms of text representations for use as features.

Zhang *et al.* (2016) proposed the use of an adapted CNN – Pronunciation-based CNN (PCNN) – which used phoneme codes as input. A phoneme is a sound that is perceived to have the same function by speakers of a language or dialect. In this way, the model was able to compensate for misspelt words since the phoneme codes for both the correct and misspelt words will be similar, thereby reducing the noise and errors commonly present in social media data. The merits of their approach were demonstrated in its improved performance compared to two other CNN models and traditional ML models. Zhao *et al.* (2020) shared similar motivations (i.e., reducing noise introduced by misspelt words) in their work experimenting with an alternative word embedding method called Locality Sensitive Hashing-Based Word Embedding (LSHWE). Their method ensured that the word representations for misspelt words are very similar to the correct spellings. For example, both *f*ck* and *fcukk* will have similar representations allowing models to process their occurrence identically. When used as features to a number of traditional ML and DNN models, LSHWE performed better than other word embeddings representations like word2vec, GloVe and SSWE. The inherently noisy nature of social media data has been previously acknowledged (Zhang *et al.*, 2016) and has become something of a common issue faced by many researchers; hence, techniques such as these that are aimed at reducing this noise can significantly improve the quality of prediction.

In their work, Chen *et al.* (2018) identified the short-form nature of tweets and social media posts as a hindrance to the performance of word embeddings-based models as the context relied on by these models might not be easily discerned in single tweets. To mitigate this, they proposed the use of 2-dimensional TF-IDF features, and they demonstrated that this outperforms models based on pre-trained word2vec vectors.

The use of CNN was not limited to detecting textual cyberbullying only; Kumari *et al.* (2019) used text and images as inputs to a CNN model to detect cyberbullying situations where a bullying message is accompanied by an image and vice versa. Hosseinmardi *et al.* (2016) had previously experimented with image-based features for cyberbullying detection but concluded that their contribution to the detection process was insignificant. Singh *et al.* (2017) also included image-based features such as the presence of text signs, abstract-rectangles and outdoor scenery in pictures and discovered an improvement when the image-based features were combined with text features but that the use of image-based features alone degrades performance (compared to text features). Kumari *et al.* (2019) and Paul and Saha (2020) adopted a different approach in their use of image-based features

to that of Hosseinmardi *et al.* (2016) and Singh *et al.* (2017). Using the Vine⁷-based dataset created by Rafiq *et al.* (2015), Paul and Saha (2020) extracted visual features from videos by decomposing the videos into frames and extracting image features from the individual frames. These were combined with text features and used as input to a hybrid BiLSTM-CNN model which achieved better performance compared to models using text or visual features only. In the study by Kumari *et al.* (2019), both image and text were separately converted into two-dimensional matrices and then merged into a single matrix that served as the input to the neural network. They achieved their best results using a single layer CNN compared to multi-layer models. As the experiments did not include control experiments using text features only, it is difficult to determine the contributions of the images to the detection process, but it nevertheless demonstrated the possibility of using actual images as inputs for cyberbullying detection and, along with the work of Paul and Saha (2020), both studies could herald more research in the use of multi-media features for cyberbullying detection.

A Recurrent Neural Network (RNN) is another deep neural network architecture that has been successfully applied to cyberbullying and online abuse detection (Chandra *et al.*, 2018; Zhang *et al.*, 2019). RNNs are multilayer neural networks that maintain a vector of hidden activations that are propagated through time (Bai *et al.*, 2018) with each layer representing observations at specific times. They differ from CNN, which are feed-forward neural networks (i.e., connections between the nodes do not form a cycle), as the connection between the nodes forms a directed graph. One way to think about an RNN model is that it can store information about what has been calculated so far. RNNs, however, do have a short-memory problem which means they can forget previously processed information in longer sequences (e.g., a paragraph of text).

To address this short-memory problem, RNN-variants like LSTM (Long Short-Term Memory) (Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005) and GRU (Gated Recurrent Unit) (Cho *et al.*, 2014) were invented. These use mechanisms called gates to regulate information flow, allowing them to decide which information is vital and what to discard. GRU is computationally less expensive than LSTM, and its ability to process long text sequences makes it very attractive as an NLP technique. Both LSTM and GRU have been used for cyberbullying detection in recent works such as that of Cheng *et al.* (2019); Saha and Senapati; Ibrahim *et al.* (2019); Anindyati *et al.* (2019); Niu

⁷vine.co

et al. (2020) and Raisi and Huang (2018a) with impressive results. Agrawal and Awekar (2018) experimented with different LSTM models (along with a CNN) model to detect cyberbullying on three different datasets sourced from Formspring, Twitter and Wikipedia and surpassed previous state-of-the-art results achieved on these datasets by traditional ML models. Their work was significant as it was the first to attempt to transfer the learning of DNN models across domains; while the results of their experiments were mixed, they crucially demonstrated the possibility of transferring knowledge gained by a DNN model on one social media platform to another. The same experiments were repeated and validated to an extent by Dadvar and Eckert (2020); notably, they used the same DNN models on another dataset sourced from YouTube, and the DNN models outperformed traditional ML models that were trained on the YouTube dataset. The implication is that in addition to improving on the state-of-the-art results achieved by traditional ML, DNN models can generalise better on different types of data, which is essential to making cyberbullying detection models available in tools accessible to the public. Furthermore, unlike traditional ML models, deep learning models do not require the extensive feature engineering associated with traditional ML algorithms, thus saving valuable time and effort in the training process.

2.3.3 Transformer-Based Cyberbullying Detection

In 2017, a new neural network architecture called Transformer was released by Vaswani *et al.* (2017) ushering in an era of Transformer-based advances in natural language processing. A Transformer is a neural network model that improves on the foundations laid by earlier deep neural networks models like the GRU and LSTM discussed in the previous section. Like RNNs, they are designed to handle large sequential data (like a paragraph of text) but, unlike RNNs, Transformers do not require that the sequential data is processed in order. Transformers examine the data passed into the network and make decisions about the importance of various parts of the data. Important parts are kept and contribute to the learning, while unimportant parts are ignored and discarded. The process by which a Transformer performs this selective learning is called an attention mechanism. The use of an attention mechanism is not novel; it has been successfully used in RNNs and LSTMs to improve performance (Luong *et al.*, 2015). Transformers, however, dispense with recurrent and convolutional networks and rely solely on attention

mechanisms as the primary means to facilitate learning (Vaswani et al., 2017). FIGURE 2.2 illustrates a Transformer model architecture.

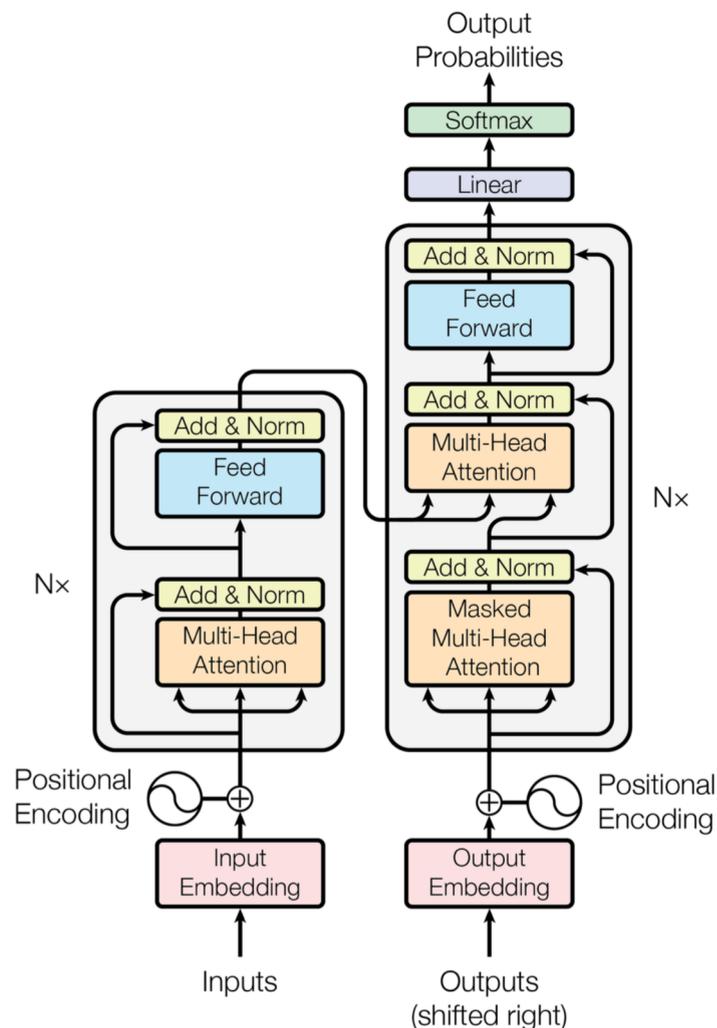


FIGURE 2.2: The Transformer Model Architecture. Source: Vaswani et al., 2017.

The left side of FIGURE 2.2 is called the Encoder and is a network layer that takes an input sequence (e.g., an English sentence) and maps it into a higher dimensional vector which is then fed into another network layer (the Decoder – on the right side of the diagram) that converts it to an output sequence. The output sequence can be the sentence in another language if the task being performed is a language translation task or values representing attributes of the sentence as per cyberbullying detection. Pradhan *et al.* (2020) repeated the experiments conducted by Agrawal and Awekar (2018) using a Transformer model in place of the CNN, and LSTM variant models used in the original experiments and the Transformer model outperformed the results achieved by the DNN

models in the original experiment. This suggests that the performance improvement witnessed in many NLP tasks by Transformer-based models can be replicated in cyberbullying detection. Within the last year, research in detecting cyberbullying and offensive language has almost exclusively used Transformer-based models. This attests to their superior performance in various NLP tasks, including cyberbullying detection compared to other machine learning models. Rather than using the original Transformer model as performed by Pradhan *et al.* (2020), the norm for researchers now is to utilise Transformer-based models that already ‘understand’ the language of the target task due to being pre-trained on a large text corpus written in the target language.

BERT (Bidirectional Encoder Representations from Transformers) is an example of such pre-trained models. It is a multi-layer bidirectional language representation model based on the Transformer (Devlin *et al.*, 2018). A language representation model (discussed in Section 2.3.2) is a way to represent natural language so it can be digested by ML models. Traditionally, machine learning methods predict a word token in a sequence based on the n tokens before or after the word. BERT, however, utilised different training strategies, namely Masked Language Modelling (MLM) and Next Sentence Prediction (NSP), and achieved better state-of-the-art results on many standard NLP tasks compared to other models at the time. In Masked Language Modelling, the model is trained to predict a word based on the tokens (i.e. words) before and after it. This is done by selecting a random sample of tokens in the input sequence and replacing them with a special mask token. The objective of MLM is to therefore calculate the cross-entropy loss on predicting the masked tokens.

NSP involves training the model to learn the relationship between a pair of sentences. Positive training data is created by extracting consecutive sentences from the same document, while negative samples are created by sentence pairs constructed from different documents. The training objective in NSP is to calculate the binary classification loss for predicting if two sentences follow each other in the original text. The BERT framework provided two modes of operations, namely pre-training and fine-tuning. Pre-training allows the model to be trained on unlabelled data to fundamentally understand a language. A 3.3 billion-word corpus sourced from the BooksCorpus (Zhu *et al.*, 2015) and Wikipedia was used to pre-train BERT to ‘teach’ it the English language. Fine-tuning is a supervised learning process whereby the pre-trained model is further

trained on domain-specific data to perform downstream tasks such as question answering, sentence completion and text classification.

The success of BERT has led to the development of BERT-like models such as DistilBERT, XLNet and RoBERTa, which were all designed to improve on BERT. DistilBERT (Distilled BERT) is a compacted BERT-based model (Sanh *et al.*, 2019) that requires fewer computing resources and training time than BERT due to using about 40% fewer parameters than BERT. DistilBERT preserves most of BERT's performance gains with only about 3% loss in language understanding compared to BERT. Similarly, AIBERT (Lan *et al.*, 2019) used parameter-reduction techniques to improve the original BERT's memory consumption and training speed while improving on the state-of-the-art results achieved by BERT on many NLP benchmark test. XLNet (Yang *et al.*, 2019) is an autoregressive BERT-like model designed to overcome some of BERT's limitations. XLNet was pre-trained on over 32 billion-words and achieved a 2%-15% improvement on the results attained by BERT on various standard NLP tasks.

RoBERTa (Robustly Optimized BERT pretraining Approach) is an optimised BERT-based model (Liu *et al.*, 2019) that improves on BERT's performance via four key adjustments to the training process. The first is using a dynamic masking strategy that generates a new masking pattern every time a training sequence is fed to the model. In comparison, BERT's static masking strategy results in the same mask being used four times on each text sequence during training. Another adjustment made compared to BERT was removing the NSP Loss during training. The NSP Loss is an auxiliary measure that indicates if sentence pairs are from the same document or different ones. The developers of RoBERTa discovered that the removal of this NSP Loss resulted in an improvement in downstream task performance. BERT uses a character-level Bytes-Pair Encoding (BPE) (Sennrich *et al.*, 2016) scheme to generate text representations. BPE represents word tokens by substituting common pairs of characters by another that is not present in the token. For example, the pair "aa" in "faaaat" can be represented by "Z" such that "faaaat" becomes "fZZt". Instead of using a character-level encoding, RoBERTa utilises a bytes-level strategy that substitutes consecutive bytes of data with a byte that does not occur in the data resulting in millions of additional parameters for RoBERTa. Finally, RoBERTa was pre-trained on substantially more data (161GB versus BERT's 16GB) using larger data batches than BERT, allowing for a more efficient training process

through parallelisation. The resultant effect of these modifications is an up to 20% performance improvement on standard NLP state-of-the-art tests compared to BERT.

BERT appears to be exceedingly popular amongst Transformer-based models for use in cyberbullying and offensive language detection as seen in recent works like that of Malte and Ratadiya (2019); Sohn and Lee (2019); Liu *et al.* (2019); Mozafari *et al.* (2020); Yadav *et al.* (2020); Paul and Saha (2020); Tanase *et al.* (2020); Elmadany *et al.* (2020) and Dadas *et al.* (2020). It has been successfully used in detecting online abuse in different languages including English (Malte and Ratadiya, 2019; Yadav *et al.*, 2020; Liu *et al.*, 2019), Hindi (Malte and Ratadiya, 2019), Mexican-Spanish (Tanase *et al.*, 2020; Guzman-Silverio *et al.*, 2020), Chinese (Sohn and Lee, 2019), Polish (Dadas *et al.*, 2020) and Arabic (Elmadany *et al.*, 2020) and, in all instances, it improved on the results achieved by traditional and DNN models. While the pre-trained version of BERT is more frequently used, fine-tuning the model has also been explored (Sohn and Lee, 2019; Elmadany *et al.*, 2020; Mozafari *et al.*, 2020) to equally impressive results. Aside from BERT, other Transformer-based models used in detecting online abuse include DistilBERT and ALBERT as seen in the work of Zinovyeva *et al.* (2020) Tripathy *et al.* (2020).

Since its introduction in 2017, the Transformer has progressed the state-of-the-art in many NLP tasks, and it is therefore not surprising that its popularity has extended into the field of cyberbullying and online abuse. As recent innovations in machine learning such as massively deep language models like Turing NLG⁸ and GPT-3⁹ continue to push the boundaries of NLP, it is not inconceivable that the boundaries of cyberbullying detection research could be similarly pushed. Existing models like BERT and RoBERTa have already demonstrated the high performance achievable by deep language models, and the massive deep language models promise even better performance. The challenge for researchers is to think beyond experimental results to developing practical tools that make use of these models to provide viable and accessible solutions to cyberbullying and online abuse.

⁸microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/

⁹openai.com/blog/openai-api/

2.4 Systems, Tools and Applications for Preventing Cyberbullying

The mitigation and prevention of cyberbullying using technology can be approached from two perspectives, namely:

- i. Raising awareness and encouraging positive emotions such as empathy amongst users; and
- ii. Detecting online abuse and mitigating its effect using punitive actions such as blocking, reporting and deleting abusive messages and senders.

In the former category, studies like those of Mancilla-Caceres *et al.* (2012, 2015); Calvo-Morata *et al.* (2018); Garaigordobil and Martínez-Valderrey (2018) and DeSmet *et al.* (2018a) adopted a serious games approach to combating bullying. Serious games are games intended for purposes other than pure entertainment. They are educational tools designed to impart learning that can be transferred into the real world and as cyberbullying prevention tools; they adopt a reflective approach by making users aware of the impact of inappropriate behaviour.

Conectado (Calvo-Morata *et al.*, 2018) is a serious game designed to increase students' cyberbullying awareness. The game places students in the role of bullying victims and encourages them to reflect on how this makes them feel and develop strategies for managing the process. A game like this can be an effective tool for increasing empathy amongst children and help reinforce a moral core, especially amongst young children. The Friendly ATTAC (Adaptive Technological Tools Against Cyberbullying) game developed by DeSmet *et al.* (2018a) is similarly aimed at encouraging positive attitudes amongst young people; it places them in a made-up scenario involving online abuse with players expected to make decisions at key points in the game. When the correct choice is selected from a list of options, the player is rewarded with a pleasant buzz, and poor decisions are accompanied by a negative buzz. Likewise, Cybereduca (Garaigordobil and Martínez-Valderrey, 2018) is a trivia pursuit game designed to raise cyberbullying awareness amongst young people and encourage positive behaviours.

CyberBullet (Mikka-Muntuomo *et al.*, 2018), Cyberhero Mobile Safety (Hswen *et al.*, 2014) and SimSafety (Cebolledo and Troyer, 2015) simulates a fictitious Online Social Network (OSN) and is aimed at teaching teenagers how to recognise cyberbullying situations and deal with such situations. Serious games are just one way to positively influence young people’s attitudes to cyberbullying. Other works in this category include that of Dinakar *et al.* (2012) who advocated the use of reflective interfaces to make apparent the impact of offensive remarks when communicating online and Fan *et al.* (2016) proposal of a new social media platform (SMP) designed from the onset to discourage cyberbullying. Social media has essentially become a platform for people to showcase their living experiences and an intrinsic part of their identity (Gündüz, 2017). As discovered by Kumar *et al.* (2011), social media users exhibit a herd mentality when migrating to new platforms. Users are more likely to migrate to a new online social network after a substantial number of their friends on an existing network have moved to the new platform. As such, a proposal such as that of Fan *et al.* (2016) would likely face an adoption barrier in attracting significant numbers of users from major OSNs like Facebook, Instagram, Twitter and TikTok to the new SMP. The use of a reflective user interface as suggested by Dinakar *et al.* (2012) is a more practical solution that can be implemented into existing platforms as well as new tools. The effectiveness of reflective user interfaces can be seen on popular SMPs like Facebook, Instagram and Twitter where they are used to hide sensitive content from users by default (see FIGURE 2.3) or identify tweets glorifying violence (FIGURE 2.4). A natural extension for this feature would therefore be to display similar warnings for other forms of online abuse such as cyberbullying.

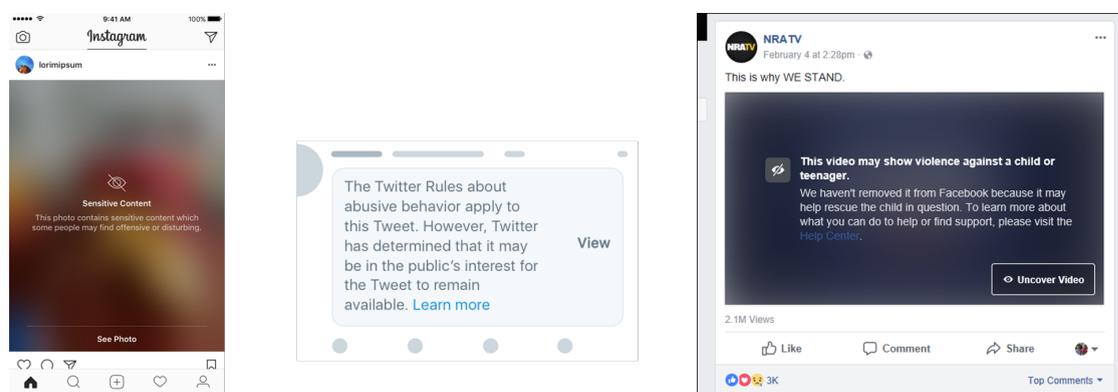


FIGURE 2.3: Use of reflective interface on Instagram, Twitter and Facebook to warn users of sensitive content.

Studies that favoured the punitive actions approach included the work by Lempa *et al.* (2015) that developed a ‘sentence checker’ mobile app to check text messages for



FIGURE 2.4: Example of a tweet identified as glorifying violence by Twitter.

offensive content before sending. The app, however, cannot be used to send messages and is not integrated with other messaging applications; as such, users have to re-type out the checked message to send via a messaging application. This introduces usability and efficiency concerns and is highly likely to serve as an adoption barrier for the app. The McDefender app developed by Vishwamitra *et al.* (2017) improves on attempts such as that of Lempa *et al.* (2015) by integrating to the Facebook app. Their McDefender app 'listens' to keystrokes directed at the Facebook app and analyses the words entered in the Facebook app for offensive content. The McDefender app can then initiate actions such as displaying a warning to the user, blocking the Facebook app or even alerting an adult if cyberbullying is detected. While this is an improvement on the approach adopted by Lempa *et al.* (2015), the app's implementation introduces a number of concerns. By listening for keystrokes entered in the Facebook app, McDefender operates in a similar manner to spyware which may cause reputational issues for the app and discourage potential users. It is also doubtful that an app operating in this manner would be accepted in major app stores. Additionally, as the app can only detect keystrokes entered within the Facebook app, it can be easily thwarted by accessing Facebook via a web browser or using one of the many 'unofficial' Facebook mobile apps. Moreover, the app would have to be updated in tandem with the Facebook app, and if a future update to the Facebook app changes its internal identifier (which is used by McDefender to identify which app to

monitor), the McDefender app will no longer be able to detect text entered in the Facebook app.

Weider *et al.* (2016) designed a messaging app with an inbuilt cyberbullying detector as an alternative to messaging tools like WhatsApp, Snapchat and Telegram. This, however, places the app in direct competition with established messaging platforms with millions of existing users, thus affecting the app's likely uptake amongst potential users. Social Net (Shome *et al.*, 2019), an OSN designed to discourage cyberbullying and promote positive online behaviours, would likely face similar adoption concerns. As a cyberbullying prevention approach, the creation of new systems and tools to compete with existing and popular SMPs is always likely to encounter difficulty enticing users. Users already have deep connections (via their contacts and friends) to existing social media platforms, and for a user to consider adopting a new online SMP, a significant proportion of their existing social network must be present on the new platform. Integrating cyberbullying prevention tools to existing platforms is, therefore, a better strategy as the adoption barrier is significantly lower in this case.

Furthermore, a new platform will have to provide a fundamentally different social proposition to what is currently available on existing platforms, and the potential to reduce instances of online abuse (which affects a minority) is unlikely to sway the majority of users in this regard. There could also be potential privacy concerns with Social Net as the platform includes a component that continuously takes pictures of users while on the social network. These facial expressions are then analysed to determine the users' emotional state, and this is used to predict if they are engaging in abusive behaviour. It is difficult to imagine users being comfortable with this 'big brother' like monitoring for an application designed for social networking, regardless of the reasoning behind this. While the system achieved impressive results in determining when users are engaged in abusive behaviour from the capture of facial expressions, the authors did not perform controlled experiments with other types of features (e.g., text features) for comparison.

Talukder and Carbanar (2018) AbuSniff and Silva *et al.* (2016, 2018) BullyBlocker are other cyberbullying prevention tools designed to work with Facebook. Both improved on McDefender's approach by integrating to the Facebook API instead of the app itself. AbuSniff was designed to identify potentially abusive contacts from a user's friends list. It makes such judgements based on users' responses to a series of questions about their

friends and social media usage. While integrating to the Facebook API is undoubtedly an improvement over McDefender, expecting users to answer the same set of questions for every contact on their friends lists (which can include hundreds of contacts) is impractical. An approach like that of BullyBlocker (Silva *et al.*, 2018) is perhaps a preferable option. The system extracts and analyses data from a Facebook account and uses this information to compute a cyberbullying risk score for the user. Similarly, CyberDect (López-Martínez *et al.*, 2019) uses the Twitter API to retrieve tweets for specified users and their followers and analyses the tweets to decide if they are being bullied or engaging in abusive behaviour. As such, it is less of a cyberbullying prevention tool than an analysis tool that can potentially be used by human moderators to investigate bullying reports. In this regards, it shares similarity with the Cyberbullying Response System proposed by Oh (2019), a system that can retrieve data about users from multiple online social networks and determine if they are engaging in abusive behaviours.

The most significant limitation of these studies, however, is that of the availability of the tools developed. None of the tools and applications described in the studies above is available to the public for use, highlighting a need for researchers to think beyond the process of detection to the viability and practicability of making their work available for use. An overview of these studies is presented in TABLE 2.2

Study	Approach		Implementation		
	Reflection	Punitive	Mobile	Serious Games	API
Chatzidaki et al. (2011)	✓			✓	
Dinakar et al. (2012)	✓				
Mancilla-Caceres et al., (2012; 2015)	✓			✓	
Hswen et al. (2014)	✓			✓	
Lempa et al. (2015)		✓	✓		
Cebolledo and Troyer (2015)	✓			✓	
Fan et al. (2016)	✓				
Weider et al. (2016)	✓		✓		
Vishwamitra et al. (2017)		✓	✓		
Calvo-Morata et al. (2018)	✓			✓	
Garaigordobil and Martínez-Valderrey (2018)	✓			✓	
DeSmet et al. (2018b.)	✓			✓	
Mikka-Muntuumo et al. (2018)	✓			✓	
Lazarinis et al. (2019)	✓			✓	
Shome et al. (2019)	✓				
López-Martínez et al. (2019)		✓			✓
Oh (2019)		✓			✓

TABLE 2.2: Automated cyberbullying prevention studies reviewed.

2.5 Mobile Apps for Cyberbullying Prevention and Mitigation

In recent years, mobile applications have become central to the use of social media. It is estimated that 42% of global social media usage is via a mobile device, with regions like East Asia, America and Europe demonstrating even higher rates of 70%, 61% and 59%, respectively (Statista, 2019). In the UK, 86% of the time spent online by people aged 13+ years is via mobile apps (Ofcom Research, 2019). This popularity of mobile apps also extends into the field of cyberbullying prevention and a simple search for the word

“cyberbullying” on the major mobile app stores reveals a varied selection of software applications, games and educational tools related to cyberbullying prevention and mitigation. With there being so many apps on the apps stores purporting to prevent cyberbullying, validating the veracity of these claims is of utmost importance to this PhD research. A survey was, therefore, conducted to discover and review existing mobile applications that can assist with cyberbullying and are available to the public to assist cyberbullying prevention.

2.5.1 Method

The main objective of the survey was to explore the mobile apps that have been developed to detect and prevent cyberbullying and online abuse, the functionalities provided by these applications, and their effectiveness in typical cyberbullying scenarios. The survey was, therefore aimed at answering the following questions:

Q1: What mobile apps are available to combat cyberbullying?

Q2: What are the approaches used by these apps to tackle cyberbullying?

Q3: How effective are these apps in their approaches?

The survey method comprised 3 phases which were:

- (i) initial search for cyberbullying prevention apps;
- (ii) filtering and shortlisting of search results; and
- (iii) evaluation of the apps.

The activities conducted within each phase are discussed below and illustrated as a flow chart in FIGURE 2.5.

Phase 1: Initial search for cyberbullying prevention apps

Searches using the keywords “bully”, “cyberbully” and “cyber-bully” were conducted on the Google Play and Apple App stores in July 2020 to discover cyberbullying prevention and mitigation mobile apps. An initial list comprising 239 apps across both app stores was compiled after eliminating duplicate entries from the search results. Apps that do not provide automated features that can help detect and prevent online abuse and

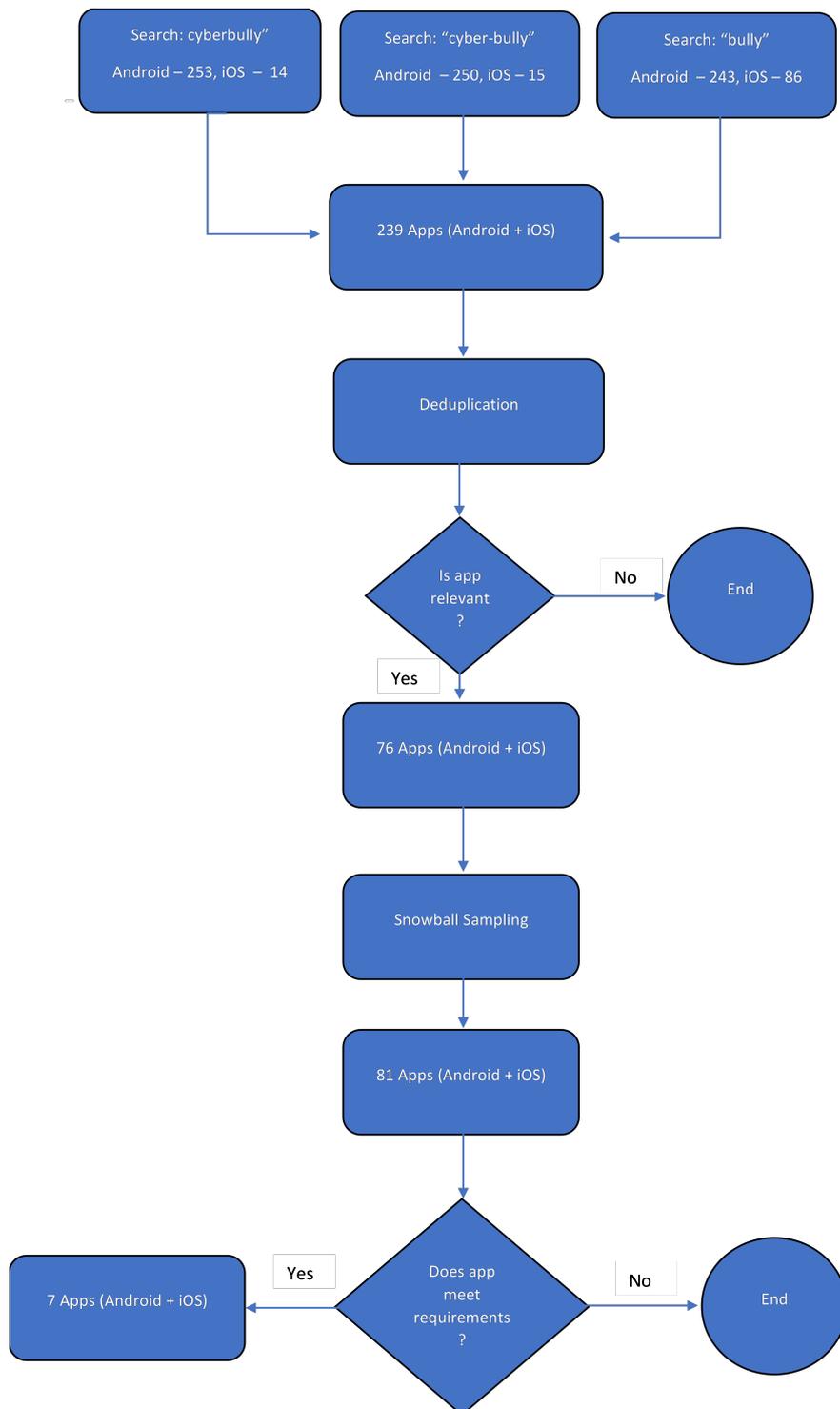


FIGURE 2.5: Mobile apps survey search and selection process flow chart.

cyberbullying (e.g., entertainment games and information-only apps) were removed from the list resulting in a shortlist of 76 apps. The 76 apps were then used in a snowballing technique to discover more qualifying apps. This was done by reviewing the apps included in the “*Similar Apps*” (Google Play Store) and “*You May Also Like*” (Apple App Store) lists (see Figures 2.6 and 2.7) on the app page for each one of the 76 apps. For each of these apps, any previously undiscovered app featured in the “*Similar Apps*” and “*You May Also Like*” list that met the selection criteria was included, resulting in an expanded list of 81 apps.

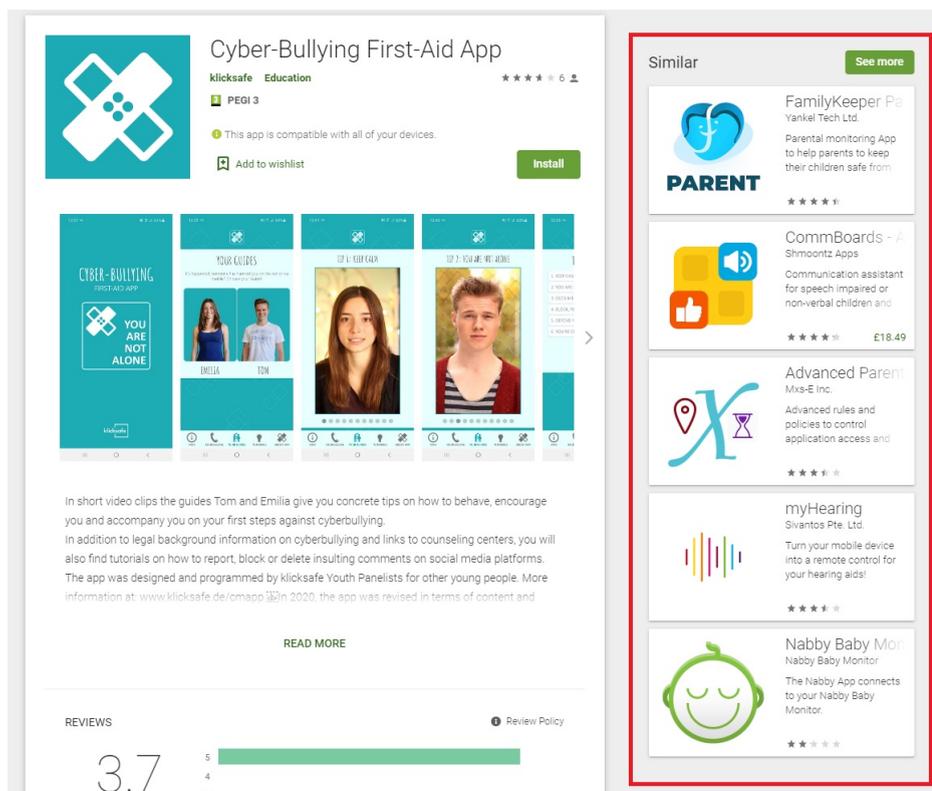


FIGURE 2.6: ‘*Similar Apps*’ listing on an app’s page on the Google Play Store.

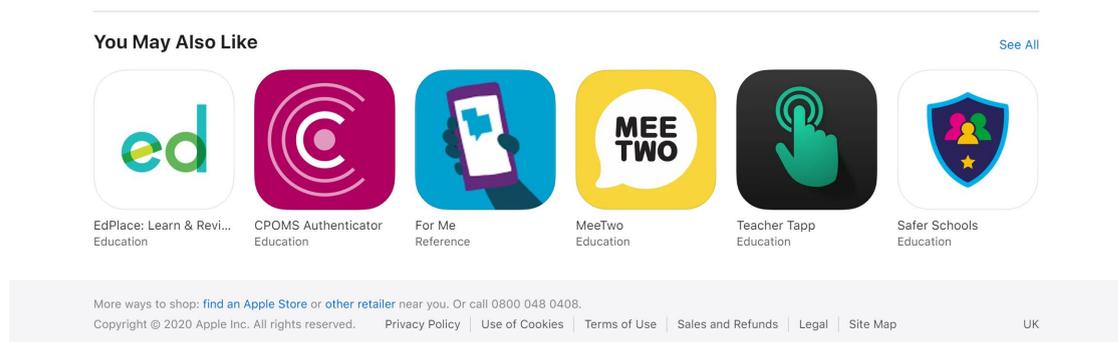


FIGURE 2.7: ‘*You May Also Like*’ apps listing on an app’s page on the Apple App Store.

Phase 2: Filtering and shortlisting of search results

Of the 81 mobile apps shortlisted, 79 of them can be categorised as parental monitoring applications. These are apps that allow parents to monitor children's activities on their mobile phones (in the same manner as spyware) and enforce device usage restrictions. They are usually implemented as a pair of linked apps: a parent app with a monitoring dashboard and a companion app installed on the child's mobile phone. The companion app sends detailed usage statistics to the parent's app and enforces restrictions set by the parent on the child's phone. This is achieved using the following tactics:

- (i) preventing access to inappropriate websites (e.g., pornography, gambling);
- (ii) preventing access to websites blacklisted by parents;
- (iii) preventing access to mobile apps blacklisted by parents;
- (iv) enforcing usage restriction set by parents;
- (v) detecting inappropriate and risky behaviours on social media; and
- (vi) detecting the use of inappropriate language.

While controlled and supervised Internet usage is sometimes used as a cyberbullying prevention strategy (Mazari, 2013), its efficacy as a long-term solution has been questioned (Sasson and Mesch, 2014). Additionally, the primary risk posed by inappropriate websites like pornographic and gambling websites is not that of cyberbullying; as such, preventing access to these sites is not a valid cyberbullying prevention strategy. Of the above tactics, detecting inappropriate and risky behaviours on social media (v) and detecting improper language use (vi) are the most relevant features that can assist with cyberbullying prevention and mitigation. An assessment of the mobile apps was conducted to identify the ones that meet these two requirements. This was done by reviewing all the available product literature for these apps including websites, app store descriptions, user reviews and testimonials and checking for references to these two functionalities (i.e., (v) and (vi)). If it was not possible to ascertain the presence or absence of these features in the app from the product literature, then the app was installed and a quick appraisal conducted by running the app for a short period to make

the determination. Five parental monitoring apps (Bark for Kids^{10 11}, Bosco Family Safe^{12 13}, Sentry Parental Control¹⁴, Surfie Parental Control^{15 16}, MMGuardian Parental Control^{17 18}) and two non-parental monitoring apps (BBC Own It^{19 20} and ReThink^{21 22}) met these requirements and made up a final shortlist of 7 apps.

Phase 3: Apps Evaluation

To evaluate the efficacy of these apps for cyberbullying prevention and mitigation, tests relating to the two functionalities of interest were created and applied. The tests were designed to evaluate the apps' abilities to accurately detect different types of abusive messages when sent or received on a number of dummy social media accounts created to facilitate the evaluation. Eight Gmail accounts were created for this purpose with five accounts designated as 'child' accounts and the other three as 'parent' accounts. The child accounts were used to create dummy profiles on Twitter, Instagram and Facebook, which were then used to conduct the tests. The three 'parent' accounts were used to monitor the 'child' accounts' social media activities via the parental monitoring applications. Details of all eight accounts are provided in TABLE 2.3.

Two Android smartphones and two iPhones were used to conduct the tests with one phone per mobile Operating System (OS) designated as the parent phone and the remaining two phones (one per OS) designated the child phones. The accompanying parent apps for the parental monitoring applications were installed on the parent phones while the child variants of the apps were installed on the child phones. All phones were assigned working mobile numbers with activated data plans so that SMS and WhatsApp could also be used to send messages. The tests devised to evaluate the apps were based on the apps' functionalities advertised in the product literature and are therefore not exhaustive of what is possible with regards to cyberbullying prevention. The tests involved sending messages containing apparent and subtle offensive content (see

¹⁰apps.apple.com/us/app/bark-connect/id1477619146?ls=1

¹¹play.google.com/store/apps/details?id=cm.pt.barkparent

¹²apps.apple.com/il/app/bosco-family-safety-locator/id1169993252

¹³play.google.com/store/apps/details?id=com.bosco.boscoApp

¹⁴play.google.com/store/apps/details?id=com.sentry.parental

¹⁵apps.apple.com/gb/app/surfie-parent/id997309073

¹⁶play.google.com/store/apps/details?id=com.puresight.surfie.parentapp

¹⁷apps.apple.com/app/apple-store/id951476346

¹⁸play.google.com/store/apps/details?id=com.mmguardian.parentapp

¹⁹apps.apple.com/gb/app/bbc-own-it-keyboard-and-diary/id1444459647

²⁰play.google.com/store/apps/details?id=uk.co.bbc.ownit

²¹apps.apple.com/us/app/rethink-stop-cyberbullying/id1035161775

²²play.google.com/store/apps/details?id=com.rethink.app.rethinkkeyboard

Account Identifier	Role	Online Social Network			Description
		X	X	X	
Bullstoptest1@gmail.com	Explicit Cyberbully	✓	✓	✓	A dummy online abuser account used to send explicitly offensive messages to the victim account and has also shared explicitly offensive content publicly
Bullstoptest2@gmail.com	Subtle Cyberbully	✓	✓	✓	A dummy online abuser account used to send subtle bullying messages to the victim account and has also shared subtle but offensive content publicly
Bullstoptest3@gmail.com	Victim	✓	✓	✓	A dummy cyberbullying victim account that receives the sample abusive messages sent by the bullying account.
Bullstoptest4@gmail.com	Neutral	✓	✓	✓	
Bullstoptest5@gmail.com	Neutral	✓	✓	✓	
Bullstoptest6@gmail.com	Parent				Dummy parent account.
Bullstoptest7@gmail.com	Parent				Dummy parent account.
Bullstoptest8@gmail.com	Parent				Dummy parent account.

TABLE 2.3: Overview of dummy accounts used to conduct the tests.

Test	Description	Functionality Being Tested	Evaluation Criteria
1	Bullstoptest1 sends ten explicit, offensive messages via Twitter, Instagram, Facebook, WhatsApps and SMS to Bullstoptest3.	Detect the presence of offensive content in the messages sent by Bullstoptest1.	The number of messages that were correctly detected.
2	Bullstoptest2 sends ten subtle but abusive messages via Twitter, Instagram, Facebook, WhatsApps and SMS to Bullstoptest3.	Detect the presence of offensive content in the messages sent by Bullstoptest2.	The number of messages that were correctly detected.
3	Bullstoptest3 receives ten explicit, offensive messages from Bullstoptest1.	Detect the presence of offensive content in the messages received by Bullstoptest3.	The number of messages that were correctly detected.
4	Bullstoptest3 receives ten subtle but abusive messages via Twitter, Instagram, Facebook, WhatsApps and SMS from Bullstoptest2.	Detect the presence of offensive content in the messages received by Bullstoptest3.	The number of messages that were correctly detected.
5	Bullstoptest4 follows Bullstoptest1 on social media.	Detect that Bullstoptest4 is interacting with a user (Bullstoptest1) that has shared explicit and offensive content publicly on Twitter, Facebook and Instagram.	The number of connection attempts to a publicly abusive user that was detected across the three social media platforms.
6	Bullstoptest5 follows Bullstoptest2 on social media.	Detect that Bullstoptest4 is interacting with a user (Bullstoptest2) that has shared subtle but offensive content publicly on Twitter, Facebook and Instagram.	The number of connection attempts to a publicly abusive user that was detected across the three social media platforms.

TABLE 2.4: Description of tests conducted to evaluate the apps.

Appendix A.1 for examples of the messages) using the dummy accounts and rating the apps on their performance on each test. The number of offensive messages or inappropriate behaviours detected was assigned as the app's score for each test. Six tests were conducted in all, and these are detailed in TABLE 2.4 with the results presented in TABLE 2.5 It should be noted that Tests 3 and 4 are the same as Tests 1 and 2 but with the app's performance evaluated from the receiver's perspective.

App	Number of messages detected (out of 10)					
	T1	T2	T3	T4	T5	T6
BBC Own It	10	4	-	-	0	0
ReThink	9	1	-	-	0	0
Bark for Kids	10	3	10	3	0	0
Bosco Family Safety	10	1	10	1	0	0
Sentry Parental Control	8	0	8	0	0	0
Surfie Parental Control	10	0	10	0	0	0
MMGuardian Parental Control	9	0	9	0	0	0

Legend	
Score	Number of offensive messages detected.
-	The feature is not available in the app.

TABLE 2.5: Results of evaluation for cyberbullying prevention apps.

The apps can be broadly divided into two categories based on the functionalities provided: parental monitoring apps and reflective keyboards. All the apps performed well in detecting the use of common offensive terms as indicated by their scores for Tests 1 and 3 but performed poorly with regards detecting subtle forms of abuse as seen in their performance on Tests 2 and 4 (the feature tested in Tests 3 and 4 are not available in the BBC Own It and ReThink apps). The apps' overall performance strongly suggests the use of wordlists to identify offensive content as they struggled to detect misspelt swear words (e.g., sluuut) and the use of non-profane terms to abuse. BBC Own It demonstrated the best performance in detecting non-profane abusive messages (e.g., "ha ha you are so fat") followed by Bark for Kids. They were the only apps that demonstrated the presence of additional logic apart from the use of wordlists.

As the only two apps that were not parental monitoring applications, BBC Own It and ReThink provided similar functionalities. Both are implemented as virtual keyboards that can be used instead of the default phone's virtual keyboard, thus providing the apps with the ability to monitor all words typed on the mobile phone regardless of the app being used. When an offensive term is detected, a message advising against the use of such terms is displayed to the user (see Figures 2.8 and 2.9). In this manner, the apps serve as reflective tools that seek to educate young people about the potential impact of the way they communicate. Both apps are capable of detecting the use of common offensive terms but, as previously mentioned, BBC Own It performed better at detecting subtler forms of abuse. It also featured additional content such as educational videos and text on mental health, general well-being and staying safe online and a daily mood tracker that can

suggest additional resources depending on the user's (self-reported) mood. In contrast, the ReThink app's only function is to serve as a reflective virtual keyboard. Between the two, BBC Own It appears the better app, it is better designed, intuitive and includes additional features not present in Rethink. Crucially, it also outperformed ReThink (and all the other apps) in the detection of offensive messages.

All five parental monitoring apps – MMGuardian Parental Control, Bark For Kids, Sentry Parental Control, Surfie Parental Control and Bosco Family Safety – performed well in detecting offensive messages containing profane words but only Bark for Kids demonstrated a performance close to that of BBC Own It in detecting subtler forms of abuse. Along with Bosco Family Safety, the two apps featured well-designed user interfaces that were easy to navigate and use. It (Bark for Kids) was the only app to perform fine-grained detection of different types of offensive content by associating the offensive messages with different labels indicating the type of offensive content detected (see FIGURE 2.10) and can be used with an extensive list of social media, messaging and email platforms (more than any of the other mobile apps). Unlike the BBC Own It and ReThink apps, it is not a free app, a trait it shares with all the other parental monitoring apps.

All five parental monitoring apps are capable of monitoring a child's social media accounts, including Twitter, Instagram, Facebook, and WhatsApp and alerting the parents when inappropriate content is detected. The key differentiator between these five apps and the several other parental monitoring apps available on the apps stores is that these apps claimed to monitor and detect risky online behaviours. In practice, their interpretation of this claim appears to be alerting parents when the time spent online by a child is over a configured threshold, or the child visits inappropriate web sites (e.g., gambling or pornographic sites). None of the apps alerted the parent when the monitored account started following a publicly abusive social media user.

At the beginning of this section, three key questions were identified, the answers to which are the objectives of this survey. The questions and the answers provided by the survey are summarised as follows:

Q1: What mobile apps are available to combat cyberbullying?

Answer: The survey uncovered seven mobile applications that stand out amongst the 81 apps included in the review. Of these, BBC Own It demonstrated the best performance

in identifying different types of offensive and bullying content and is followed by Bark For Kids in terms of the ability to detect offensive content. Both apps are intuitive and well designed.

Q2: What are the approaches used by these apps to tackle cyberbullying?

Answer: Without examining the applications' source code, it is impossible to fully confirm the approach used by the apps to detect offensive language, but the results of the tests conducted suggest the use of wordlists is common to all the applications.

Q3: How effective are these apps in their approaches?

Answer: All the apps correctly identified offensive messages containing common profane terms, but they all struggled in detecting abusive messages when profane terms are not used.

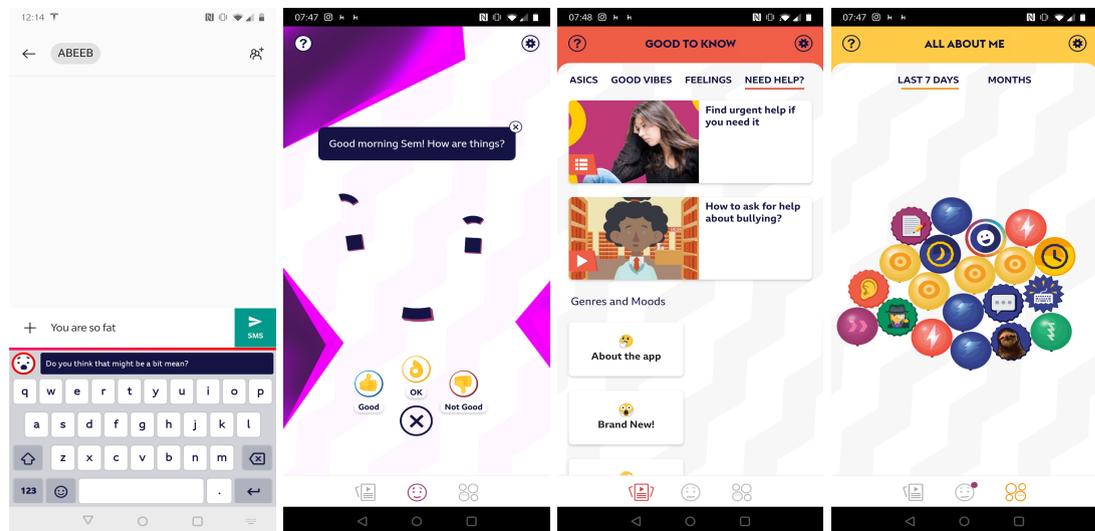


FIGURE 2.8: BBC Own It sample screens

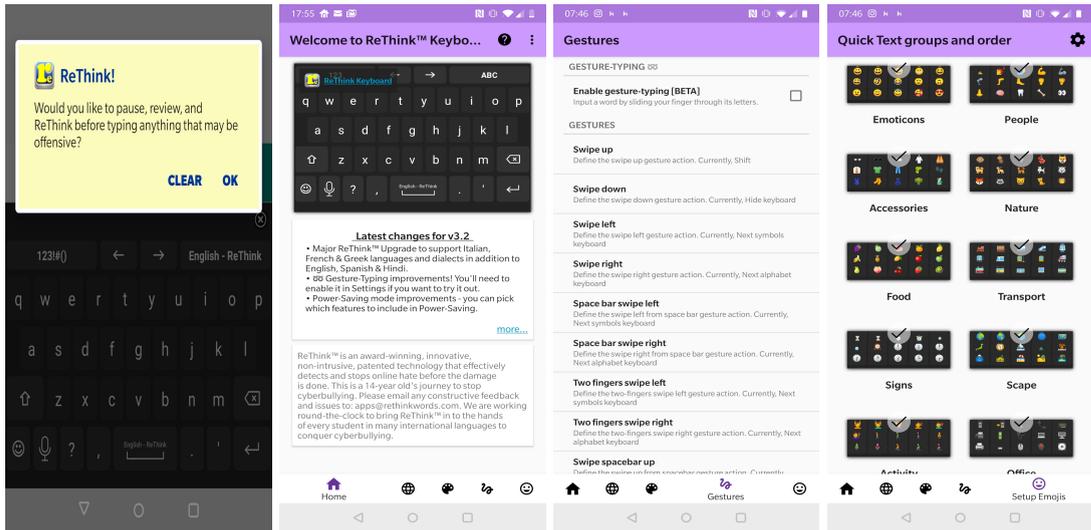


FIGURE 2.9: ReThink sample screens

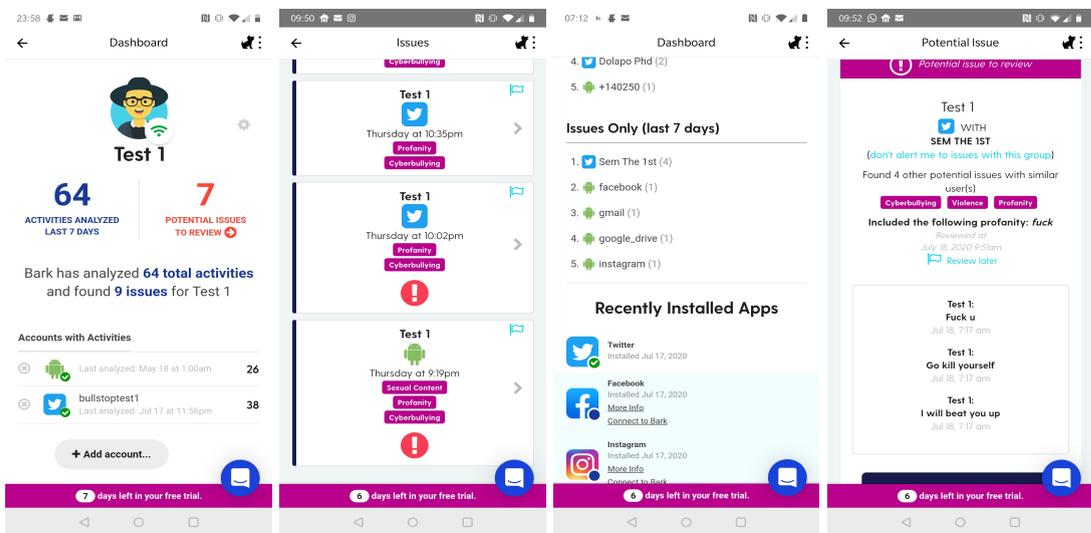


FIGURE 2.10: Bark For Kids sample screens

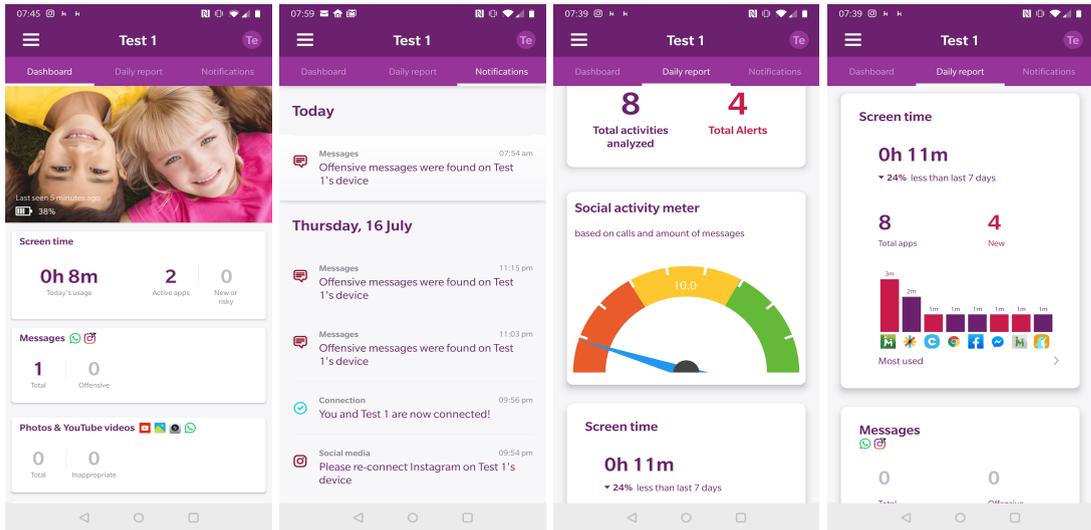


FIGURE 2.11: Bosco Family Safety sample screens

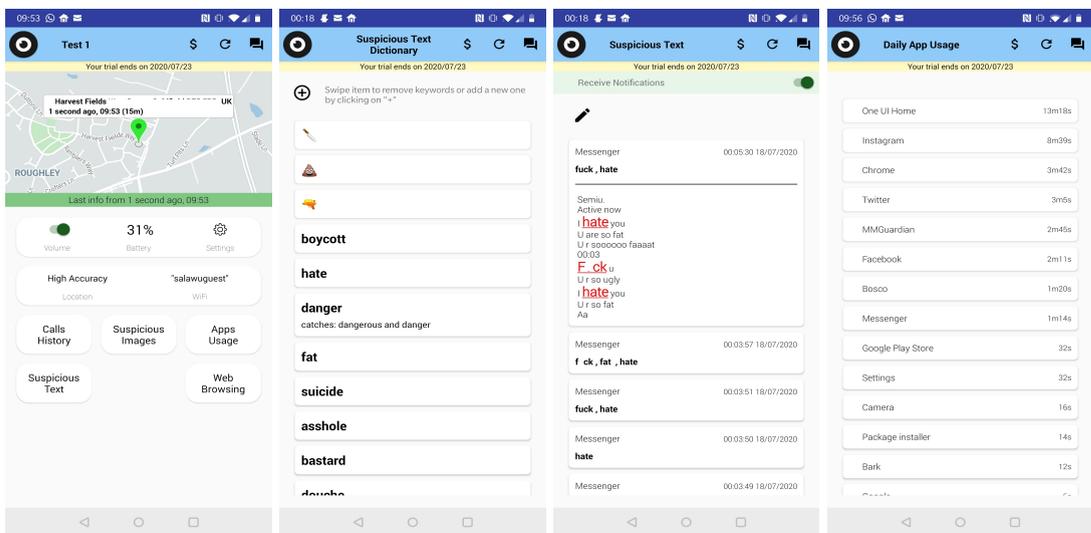


FIGURE 2.12: Sentry Parental Control sample screens

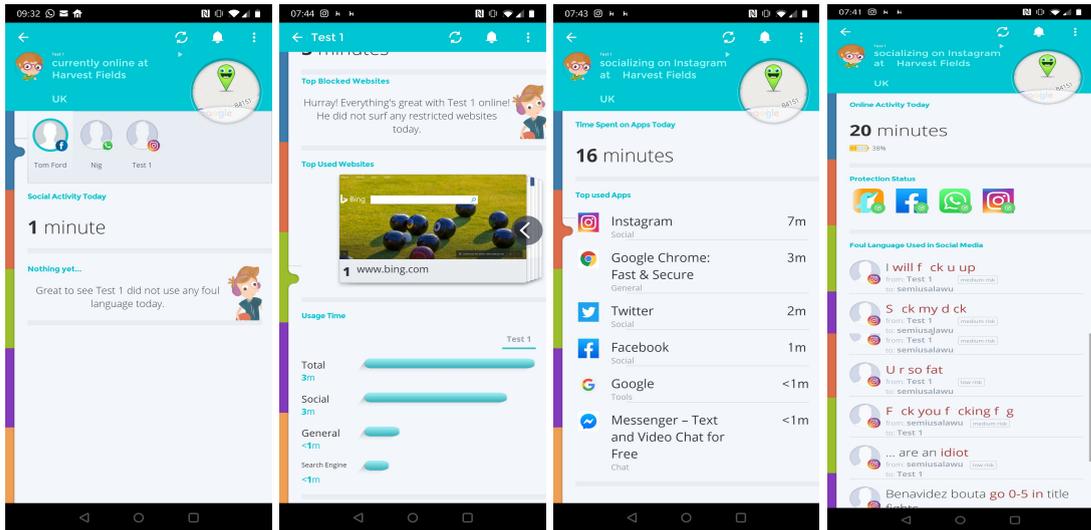


FIGURE 2.13: Surfie Parental Control sample screens

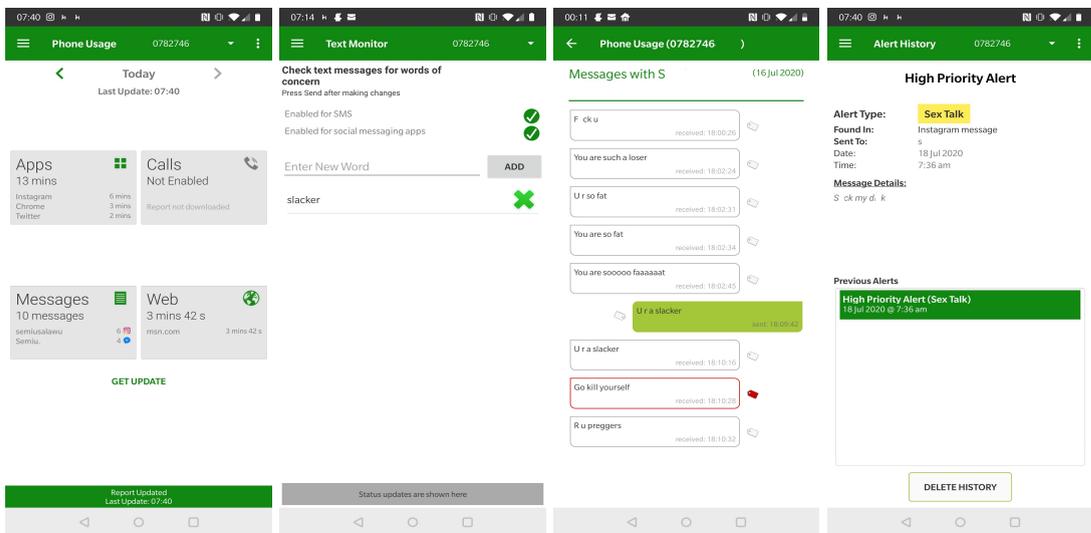


FIGURE 2.14: MMGuardian Parental Control sample screens

2.6 Cyberbullying Prevention by Social Media Platforms

As the overwhelming majority of cyberbullying and online abuse occur on online social networks, platform providers are often derided for not doing enough to stem the onslaught of cyberbullying (BBC News, 2017, 2018). In response to these accusations, SMPs have increased efforts in combating online abuse in recent years. In 2013, Facebook launched its Bullying Prevention Hub²³ to provide parents, educators, and teens with resources to

²³facebook.com/safety/bullying

deal with cyberbullying and since then it has increased efforts to discourage online abuse on the platform by removing posts promoting violence and hate speech and banning users that share such content. Similarly, Twitter introduced enhancements to its platform to help users combat online abuse; these enhancements include an online abuse help centre, filtering of inappropriate content (e.g., adult-related content are hidden by default for all Twitter users), preventing the creation of multiple accounts with the same email address or phone number, and aggressively identifying and banning accounts used to send offensive tweets (this resulted in some of the dummy accounts used in the evaluation of the mobile apps being banned for posting offensive content).

Perhaps, the most significant indication of SMPs resolve to tackle online abuse is Twitter and Instagram's trial of new features that analyse content before it is shared and then displaying reflective messages imploring the user to edit the message before posting if the post is deemed offensive (Porter, 2019; Statt, 2020). This could significantly reduce the amount of online abuse on this platform. And yet, these features have been undergoing testing for over a year and are yet to be released to the public while many other features unrelated to tackling online abuse have been released on these platforms during the same period. It is possible that the sensitive and subjective nature of online abuse necessitate additional considerations and hence the prolonged testing period. It is also likely that these platforms that are commercial entities and rely on the continuous patronage of their multitude of users to generate revenue via advertising are cautious about being accused of censoring free speech and the potential loss of users that may accompany this and thus it may become a waiting game to see which of the platforms is brave enough to implement this feature first. And while the waiting continues, so does cyberbullying.

2.7 Summary

Compared to the mature literature on traditional bullying, cyberbullying research could be considered to be in a state of vibrant adolescence. A substantial amount of research effort has been expended to understand this modern scourge of the Internet. Despite this, there are still a great many inconsistencies about cyberbullying. Researchers disagree about how it should be defined, its prevalence as reported across several studies, and to what extent its negative consequences can be genuinely attributed to it alone.

In presenting a review of the pertinent literature on cyberbullying and its automated detection, this chapter has chronicled the development of the various techniques used for the automated detection of cyberbullying. As researchers continuously seek to improve the performance of cyberbullying classifiers, they have explored a range of diverse techniques that extended from the use of traditional ML algorithms like SVM, Logistic Regression and Naïve Bayes to augmenting these algorithms with ideas borrowed from different fields like sentiment analysis, expert systems and decision support applications. These later gave way to DNN models like CNN, GRU and LSTM that demonstrated better abilities to generalise and understand written text contextually. Additionally, these DNN models are capable of transferring knowledge learned on one task or domain to another, exposing even more opportunities to improve the detection process. This and other developments have contributed to progressing cyberbullying detection research to a point where it is now possible to predict the presence of various forms of offensive content to a reasonable degree of accuracy.

With regards to the prevention and mitigation of cyberbullying using technology, two key strategies were uncovered from the literature. These are: the use of reflective techniques to prevent the occurrence of cyberbullying by raising awareness of its consequences amongst young people, and the use of punitive actions to protect victims and discourage perpetrators from re-offending. The innovative propositions discovered are, however, severely limited by a failure to fully consider the practical implications of their implementations. Consequently, these tools have not achieved the desired impact in the lives of many cyberbullying victims. Indeed very few of these tools progress beyond being research prototypes. The literature review also uncovered a disparity between the number of studies focused on improving the performance of cyberbullying detection algorithms and those directed at building tools that make use of these algorithms to prevent cyberbullying in favour of the former. This can be likened to an 'arms race' where researchers are focused on attaining better cyberbullying detection performance metrics, but less attention is devoted to creating cyberbullying prevention tools to utilise the models resulting from such research to curb the real-world damage being caused by cyberbullying.

This has thus revealed an underserved area of research, namely the creation of viable and practical tools for the prevention and mitigation of cyberbullying – tools that can achieve the impact many existing studies desired but could not achieve due to adoption

barriers created by their implementations. To achieve this level of impact requires adequate consideration for the needs and desires of the young people whose lives are the most affected by cyberbullying as well as practical considerations on the usability and technical challenges posed by the SMPs where it is most prevalent.

This dissertation, therefore, contributes to knowledge in this underserved area. It reports on the development of a mobile-based cyberbullying prevention system designed to be highly scalable and responsive to make it viable for use with modern OSN and provide users with a similar level of performance as typically experienced on popular social media mobile apps. In creating the system, the research agenda focused on developing a method that allows the use of different ML models for cyberbullying detection. This deviates from the majority of existing research in cyberbullying detection, which are focused on devising new algorithms and models to detect online abuse. By proposing an application framework that enables the tool to use the results of these other research works (and their future iterations), the research program imbued the tool with unprecedented longevity. As the state-of-the-art in cyberbullying and offensive language detection is advanced by the development of new models, the tool benefits by adopting these models to improve its performance. The system is also novel in its ability to generate a personalised online abuse classifier for each user. These personalised classifiers are initially trained using a large scale cyberbullying and offensive language dataset that uses a fine-grained annotation scheme to identify various types of cyberbullying. This dataset was purposely created by the research program to serve as the mechanism to impart relevant contextual knowledge about the different forms of cyberbullying on social media to the classifiers (see Chapter 3). Ground truth provided by end-users in the form of messages re-classified by them is then used to retrain the personalised classifiers providing them with additional insight into end-users' behaviours and communication styles.

As previously mentioned, the implementation of existing cyberbullying prevention tools may introduce adoption barriers, thus negatively impacting their acceptance amongst the intended audience. To ensure that the tool meets its potential end-users' requirements, collaborative design methodologies were used to capture the stakeholders' requirements and guide the design and technical development of the proposed novel mobile-based cyberbullying prevention application.

Chapter 3: Fine-Grained Detection of Cyberbullying on Social Media

3.1 Introduction

The standard supervised method when using machine learning techniques for cyberbullying detection is to train a model using a suitable corpus and to then predict unseen data based on the trained model. In the previous chapter, the four key tasks performed in cyberbullying detection were discussed. While all these tasks can be achieved using NLP techniques, a paucity of large, labelled datasets remains a crucial challenge for cyberbullying detection research (Dadvar and Jong, 2012). The domain-bias, composition and taxonomy of a dataset can impact the suitability of models trained on it for cyberbullying detection purposes, and therefore the choice of training data plays a significant role in the performance of these tasks. A key deliverable of this research is a novel mobile application that uses a deep learning model to identify different forms of cyberbullying and online abuse. The model's ability to predict cyberbullying instances is made possible via knowledge acquired by training it on a new English language dataset created as part of this research to facilitate the deep learning model's understanding of different types of online abuse and how these are manifested on social media.

This chapter details the methods and processes employed in sourcing, creating and annotating the new dataset and is broadly divided into four sections. Section 3.2 reviews existing cyberbullying datasets initially considered for use for this research and the reasons why they were deemed unsuitable for training the model. In Section 3.3, the dataset and the method used for creating it are discussed. Section 3.4 reports on the experiments conducted to train and evaluate different machine learning models with the

dataset to select the best performing model to use in the mobile application. The cross-domain evaluation experiments conducted to assess the dataset's suitability for the intended task in comparison to two other popular datasets are also discussed in this section. Finally, Section 3.5 concludes the chapters with a summary of the contents.

3.2 Existing Cyberbullying Datasets

As a suitable dataset was required to facilitate the training of the ML models used in the developed cyberbullying prevention system, existing cyberbullying and offensive language datasets were reviewed to identify potential candidates to use as training data for the models. The requirement was for a large labelled dataset that contained a substantial proportion of online abuse samples and utilised a labelling scheme that allowed for identifying different forms of online abuse including less frequent types such as social exclusion, threat and sarcasm.

Existing cyberbullying and offensive language datasets can be categorised into three groups based on the annotation scheme used. These are binary, multi-class and multi-label. When the aim is to simply determine if a document contains bullying content, then this is formulated as a binary classification problem with two classes (bullying and non-bullying). This approach is widely used in literature, as seen in the work of Kontostathis *et al.* (2013); Manganonkar *et al.* (2015); Rafiq *et al.* (2015) and Zhao *et al.* (2016). An observed limitation of binary-labelled datasets is the low sophistication exhibited by models trained on such datasets. This is because of the limited choice of labels available; all documents can only belong to one of two classes, restricting opportunities to explore the finer details of the identified abuse (e.g. racism or sexism). Furthermore, these types of datasets are predominantly focused on the identification of profane and aggressive text; while cyberbullying is often associated with profanity and online aggression, its highly subjective nature means that its accurate detection extends beyond the mere identification of swear words since cyberbullying can be perpetrated without the use of profane or aggressive language. Binary-labelled datasets are therefore often unable to capture this complexity. Multi-class and multi-label classification are the two other text categorisation approaches used to annotate datasets, and both improve on the limitations of binary classification.

A multi-class approach expands the number of classes to more than two, and documents are assigned to only one of the possible classes. This is the approach taken by Waseem and Hovy (2016); Chatzakou *et al.* (2017); Founta *et al.* (2018); Rezvan *et al.* (2018) and Bruwaene *et al.* (2020). The limitation with this approach is that a document can simultaneously belong to more than one class and, as a multi-class approach can only associate a document with a single class, other applicable classes are ignored, resulting in a less refined classifier. In a multi-label classification approach, documents are associated with one or more classes and, in so doing, more information about the documents is made available. This is the approach favoured by recent studies such as those by Hee *et al.* (2018) and Ousidhoum *et al.* (2019).

The Barcelona Media corpus¹, a cyberbullying and offensive language dataset sourced from a number of websites and online forums including Kongregate, Slashdot, Formspring and MySpace, featured heavily in earlier work (Dadvar and Jong, 2012; Nahar *et al.*, 2014; Huang *et al.*, 2014; Nandhini and Sheeba, 2015) on cyberbullying detection. This dataset was created over ten years ago, and many of the social networks used to source the data are now defunct. This dataset is also no longer representative of contemporary social media usage, which is now typified by social media platforms like Instagram, Twitter and Facebook. More recently, studies such as those by Chatzakou *et al.* (2017); Founta *et al.* (2018) and Davidson *et al.* (2017) have created newer cyberbullying datasets sourced from popular SMPs. As mentioned earlier, the composition of a dataset can impact its relevance to target tasks. For example, a model trained on data sourced from websites like Wikipedia and online blogs — where content is typically in article format with several hundreds and thousands of words — may not readily lend itself to use on unseen data sourced from platforms like Twitter and Instagram where content is considerably shorter (the maximum number of characters allowed in a tweet is 280) due to the possibility of the model using features learnt from the longer documents that are not present in the short documents; this difference in the documents' lengths can therefore affect a model's performance.

The distribution of classes within a dataset may also be inadequate. For example, the number of documents assigned the *bullying* label may be so small compared to those classified as *not bullying* that the model struggles to extract meaningful features to represent the minority class and may even acquire a bias in what it associates with the minority class. In such situations, oversampling techniques like SMOTE (Synthetic

¹caw2.barcelonamedia.org

Minority Over-Sampling Technique) (Chawla *et al.*, 2002) and ADASYN (Adaptive Synthetic) (He *et al.*, 2008) may be required to improve the number of minority observations within the dataset. Oversampling, however, repeats or generates new documents based on existing ones which makes the resulting dataset less natural in its textual composition.

Twitter is one of the most widely used social media platforms globally (Statista, 2020b); it is, therefore, unsurprising that it is frequently used to source cyberbullying data. Bretschneider *et al.* (2014) annotated 5,362 tweets, 220 of which were found to contain online harassment; the low proportion of offensive tweets present within the dataset (less than 0.05%), however, limits its efficacy for classifier training. More recently, studies such as those by Rajadesingan *et al.* (2015); Waseem and Hovy (2016); Davidson *et al.* (2017); Chatzakou *et al.* (2017); Hee *et al.* (2018); Founta *et al.* (2018) and Ousidhoum *et al.* (2019) have produced datasets with higher positive samples of cyberbullying and online abuse (see TABLE 3.1). Rajadesingan *et al.* (2015) labelled 91,040 tweets for sarcasm. This is noteworthy because sarcasm is rarely featured as a label in existing cyberbullying datasets even though it is regularly used to perpetrate online bullying; it remains undetected in many instances due to the difficulty in detecting sarcasm using conventional NLP techniques like sentiment analysis. As the dataset was created for sarcasm detection only, this is the only context that can be learned from the dataset. As such, any model trained with this dataset will be unable to identify other forms of cyberbullying, thus limiting its usefulness.

In creating their bi-lingual dataset sourced from ASKfm, Hee *et al.* (2018) used a comprehensive labelling scheme that acknowledges the different types of cyberbullying discovered in the retrieved post types. The dataset's effectiveness in training classifiers may, however, be affected by the low percentage of abusive documents present. The dataset created by Founta *et al.* (2018) contained a substantial number of abusive documents but suffered the same limitations mentioned above as other multi-class labelled datasets (Chatzakou *et al.*, 2017; Rezvan *et al.*, 2018; Bruwaene *et al.*, 2020). Ousidhoum *et al.* (2019) used one of the most comprehensive annotation schemes encountered in an existing dataset and additionally included a very high percentage of positive cyberbullying samples in their dataset but, regrettably, the number of English documents included in the dataset is small in comparison to other datasets and not of a sufficient proportion to ensure adequate training for English language models.

Zampieri *et al.* (2019) used a hierarchical annotation scheme that, in addition to identifying offensive tweets, also identifies if such tweets are targeted at specific individuals or groups and what the type of target it is (i.e., individual - *@username* or group – “... *all you republicans*”). Hierarchical annotation schemes have indeed shown promise as observed in their use in recent offensive language detection competitions like *hatEval*² and *OffensEval*³; that said, however, a hierarchical scheme could inadvertently filter out relevant labels depending on the choice of a first-level label and the subjectiveness of the sentence. For example, if the first-level classification objective is to determine if a post is bullying or not then a post like “*That’s worth a sh*t load of money*” would be (correctly) classified as “*not bullying*” and go no further along the hierarchy. In contrast, a flat annotation scheme will at least identify “*sh*t*” as a profane word ensuring that any applicable label is still assigned, subsequently improving the context that can be learned from the document.

Finally, Kaggle⁴, an online resource for machine learning and NLP owned by Google, provided the Kaggle Insult (Kaggle, 2012) and Kaggle Toxic Comments (Kaggle, 2018) datasets to aid offensive language detection. The Kaggle Insult dataset contains about 4,000 posts labelled using a binary annotation scheme and thus inherits the limitations observed with this type of annotation scheme. The Kaggle Toxic Comments dataset is a large corpus of almost 160,000 Wikipedia comments labelled using a multi-label annotation scheme. The dataset is, however, focused on obscene, toxic and hate-speech content and ignored other subtler forms of online abuse that are of interest to this program of research, thus limiting its suitability for the program’s purposes.

While a number of these datasets (see TABLE 3.1) possess some of the attributes desired by the research (e.g., large number of samples, fine-grained annotation scheme, high proportion of labelled offensive instances), no single dataset captured all the research program’s requirements. The dataset created by Hee *et al.* (2018) was the most appropriate dataset for the research study’s purposes but as mentioned above the total number of English documents and the percentage of this that were positive abusive samples were lower than the desired magnitude for the research program. This, therefore, necessitated the need for a novel large-scale cyberbullying and offensive

²competitions.codalab.org/competitions/19935

³sites.google.com/site/offensevalsharedtask

⁴kaggle.com

language dataset that contains a sizeable proportion of abusive content and also explored less common forms of online bullying like sarcasm, social exclusion and threats.

3.3 A Twitter-Based Dataset for Detecting Cyberbullying and Offensive Language

This section discusses the creation of a new large-scale Twitter-based dataset for detecting instances of cyberbullying and offensive language. The dataset was used to train various machine learning models as part of a series of experiments to identify the best performing model for use in the novel cyberbullying prevention mobile application.

3.3.1 Dataset Objective

The key objective for creating this novel dataset was to provide a dataset that included a substantial proportion of different forms of bullying and offensive language so that ML models can be trained to predict these types of online abuse and to do so without the need for oversampling techniques. In reviewing various samples of offensive tweets, it was discovered that a single tweet could simultaneously contain elements of abuse, bullying, hate speech, sex talk and many other forms of objectionable content. As such, attributing a single label to a tweet ignores other salient labels that can be ascribed to the tweet. Consequently, this research proposed a multi-label annotation scheme that identifies the many offensive content elements present in a single tweet. Twitter, being one of the largest online social networks with a user base in excess of 260 million (Statista, 2020b) and highly representative of current social media usage, was used to source the data.

3.3.2 Labels

As mentioned in the previous chapter, cyberbullying can be direct and indirect with direct cyberbullying typified by targeted aggression involving the use of profanity and inappropriate language. This type of cyberbullying is evident and more frequently encountered and is often the only type of cyberbullying captured in many existing datasets. In contrast, indirect cyberbullying is a subtler form of online abuse that is not

Dataset	Source	Size	% of Abusive samples	Labels	Annotation
Bretschneider et al. (2014)	Twitter	5,362	0.05	Harassment, Not	Binary
Rajadesingan et al. (2015)	Twitter	91,040	9.99	True, False	Binary
Waseem and Hovy (2016)	Twitter	16,907	31.6	Racism, Sexism, None	Multi-class
Davidson et al. (2017)	Twitter	25,296	81.5	Hate speech, offensive, neither	Multi-class
Chatzakou et al. (2017)	Twitter	9,484	40.9	Aggressor, Bully, Spammer, Normal	Multi-class
Van Hee et al. (2018)	ASKfm	English (13,698) Dutch (78,387)	English (4.73) Dutch (6.97)	Threat/blackmail, Insult, Curse/Exclusion, Defamation, Sexual Talk, Defense, Encouragement to harass	Multi-label
Rezvan et al. (2018)	Twitter	24189	12.9	Harassment, Non-Harassment, Other	Multi-class
Founta et al. (2018)	Twitter	99,800	46.1	Abusive, Hateful, Normal, Spam	Multi-class
Zampieri et al. (2019)	Twitter	14000	-	Offensive (Yes, No), Targeted (Yes, No), Target (Individual, Group, Other)	Multi-class
Ousidhoum et al. (2019)	Twitter	English (5,647) French (4,014) Arabic (3,353)	English (75.9) French (71.9) Arabic (64.3)	26 labels covering 5 attributes	Multi-label
Kaggle Insult	Various	3,947	26.6	True, False	Binary
Van Bruwaene et al. (2020)	Various	14,900	25.1	Bullying (True, False), Cyberaggression (True, False)	Multi-class
Kaggle Toxic Comments	Wikipedia	159,570	10.2	Toxic, Severe_toxic, Obscene, Threat, Insult, identity_hate	Multi-label

TABLE 3.1: Comparison of cyberbullying and offensive content datasets.

always apparent and is, therefore, more challenging to capture. As a result, existing datasets rarely contain examples of this type of cyberbullying. The annotation scheme developed was, therefore, designed to not only capture both direct and indirect bullying but also the presence of offensive and abusive language in the tweets. For example, a tweet such as “*f*ck you @username, you f*cking loser*” would be labelled as bullying, profanity and insult as it is a profane and insulting tweet with a clear intended target (i.e. the individual identified by @username) while a tweet such as “*sh*t, sh*t, sh*t*” is just a profane utterance without an obvious target and would only be labelled as profanity. The labels used to annotate the dataset are presented in TABLE 3.2.

Label	Description	Example
Bullying	Tweets directed at a person(s) intended to provoke and cause offence.	@username You are actually disgusting in these slutty pictures Your parents are probably embarrassed. . .
Insult	Tweets containing insults typically directed at or referencing specific individual(s).	@username It’s because you’re a c*nt isn’t it? Go on you are aren’t you?
Profanity	This label is assigned to any tweets containing profane words.	@username please dont become that lowkey hating ass f**king friend please dont
Sarcasm	Sarcastic tweets aimed to ridicule. These tweets may be in the form of statements, observations and declarations.	@username Trump is the most innocent man wrongly accused since O.J. Simpson. #Sarcasm
Threat	Tweets threatening violence and aggression towards individuals.	@username Let me at him. I will f*ck him up and let my cat scratch the f*ck out of him.
Exclusion	Tweets designed to cause emotional distress via social exclusion.	@username @username You must be gay huh ? Why you here ? Fag !! And I got 2 TANK YA !
Porn	Tweets that contain or advertise pornographic content	CLICK TO WATCH [link] Tinder Sl*t Heather Gets her A*s Spanks and Spreads her C*nt
Spam	Unsolicited tweets containing and advertising irrelevant content. They typically include links to other web pages	HAPPY #NationalMasturbationDay #c*m and watch me celebrate Subscribe TODAY for a free #p*ssy play video of me [link]

TABLE 3.2: Annotation scheme with examples.

3.3.3 Data Collection

As mentioned earlier, a key objective in creating the dataset was ensuring it contained a significant portion of offensive and cyberbullying samples to facilitate training without the need for oversampling; this, therefore, influenced querying strategies. Rather than indiscriminately mining Twitter feeds, a series of searches formulated to return tweets with a high probability of containing the various types of offensive content of interest was executed. For insulting and profane tweets, Twitter was queried using the 15 most frequently used profane terms on Twitter as identified by Wang *et al.* (2014). These are fuck, shit, ass, bitch, nigga, hell, whore, dick, piss, pussy, slut, puta, tit, damn, fag, cunt, cum, cock, blowjob, retard. To retrieve tweets containing sarcasm, the hashtag *#sarcasm* was used to query the Twitter Streaming and Search APIs. This strategy is based on the work of Rajadesingan *et al.* (2015) which discovered that sarcastic tweets often include *#sarcasm* and *#not* hashtags to make it evident that sarcasm was the intention

To discover prospective query terms for threatening tweets, 5,000 tweets were randomly selected from the streaming API using the 'account home' metadata as the filtering criteria. Account home identifies the location set by the user on their profile and countries classed as English-speaking were used as filters. From the 5,000 tweets, 326 tweets were classified as threatening; the hashtags used in these tweets were then retrieved and used as query terms to extract other threatening tweets. The hashtags were *#die*, *#killyou*, *#rape*, *#chink*, *#muslim*, *#FightAfterTheFight*, *#cops*, *#karen* and *#karensgonewild*. These hashtags were then used as the initial seed in a snowballing technique to discover other relevant hashtags. This was done by querying Twitter using the hashtags and inspecting the returned tweets for violence-related hashtags. The following additional hashtags were subsequently discovered through this process: *#killallblacks*; *#killallcrackers*; *#blm*; *#blacklivesmatter*; *#alllivesmatter*; *#bluelivesmatter*; *#killchinese*; *#bustyourhead*; *#fuckyouup*; *killallwhites*; *maga*; *killallniggas*; and *nigger*.

Formulating a search to retrieve tweets relating to social exclusion was challenging as typical examples were hard to come by. A similar strategy was therefore used to retrieve tweets relating to social exclusion; from the 5,000 tweets sample, six were identified as relating to social exclusion and the following hashtags retrieved from these tweets were used as query terms: *#alone*; *#dontcometomyparty*; *#idontlikeyou*; and *#stayinyourlane*. Due to the low number of tweets returned for these hashtags, we also extracted the

replies associated with returned tweets and discovered the following additional hashtags – *#notinvented* and *#thereisareasonwhy* – which were subsequently used as additional query terms. Rather than excluding retweets when querying, as is common practice amongst researchers, our process initially extracted original tweets and retweets and then selected only one if a tweet and its retweets are included in the results. This ensured relevant content was not discarded in situations where original tweets were not included in the results, but retweets were. Our final dataset contained 62,587 tweets published in the period April to June 2019.

3.3.4 Annotation Process

Language use on social media platforms like Twitter is often colloquial; this, therefore, influenced the desired annotator profile as that of an active social media user that understands the nuances of Twitter’s colloquial language use. While there is no universal definition of what constitutes an active user on an online social network, Facebook defined an active user as someone who has logged into the site and completed an action such as liking, sharing and posting within the previous 30 days (Cohen, 2015). With one in every five minutes spent online involving social media usage and an average of 39 minutes spent daily on social media in the UK (Ofcom Research, 2019), this definition is inadequate in view of the increased users’ activities on social media. An active user was therefore redefined as one that has accessed any of the major social networks (e.g., Twitter, Instagram, Facebook, Snapchat) at least twice a week and made a post/comment, like/dislike or tweet/retweet at least once in the preceding two weeks. This new definition is more in keeping with typical social media usage.

Using personal contacts, a pool of 17 annotators whose self-reported online social networking habits met our definition of an active social media user were recruited to label the dataset. Since the presence of many profane words can be automatically detected, a program was written to label the tweets for profane terms based on the 15 profane words used as query terms (fuck, shit, ass, bitch, nigga, hell, whore, dick, piss, pussy, slut, puta, tit, damn, fag, cunt, cum, cock, blowjob, retard), and the Google swear words list⁵. The profanity-labelled tweets were then provided to the annotators to alleviate this aspect of the labelling task. The annotators were assigned to ten groups of three, with each

⁵code.google.com/archive/p/badwordslist/downloads

annotator assigned to at least one group and others to more depending on their availability. The dataset was divided into batches of at least 6,000 tweets, and each annotator in a group was provided with the same set of tweets, ensuring that each tweet was labelled by three different annotators. Inter-rater agreement was measured via Krippendorff's Alpha (α) and the majority of annotators' agreement was required for each label. Krippendorff's Alpha is a reliability coefficient used in measuring the agreement amongst annotators/raters (Krippendorff, 2011). Its general form is:

$$\alpha = 1 - \frac{D_o}{D_e}$$

Where D_o is the observed disagreement among values assigned to the tweets and D_e is the disagreement if the annotations were left to chance. The value of α lies between 1 and 0 such that $\alpha = 1$ indicates perfect agreement and $\alpha = 0$ means there was no agreement between annotators which occurs when $D_o = D_e$ – i.e., the annotators' agreement is as if by chance. The Krippendorff python library⁶ was used to compute α , which was found to be 0.67.

3.3.5 Preprocessing

Preprocessing was performed on the dataset to remove irrelevant and noisy data that may hamper classifier training. As is standard for many NLP tasks, punctuation, symbols and non-ASCII characters were removed. This was followed by the removal of mentions (including a username with the @ symbol inside a tweet) and URLs. We also discovered many made-up words created by combining multiple words (e.g., goaway, itdoesntwork, gokillyourself) in the tweets. These are due to hashtags, typos and attempts by users to mitigate the characters limit imposed by Twitter. The wordsegment python library⁷ was used to separate these into individual words. The library contains an extensive list of English words and is based on Google's 1T (1 Trillion) Web corpus⁸. Lastly, the text was converted to lower case.

⁶pypi.org/project/krippendorff

⁷pypi.org/project/wordsegment/

⁸pypi.org/project/krippendorff/

3.3.6 Dataset Analysis

Profanity emerged as the dataset’s majority class, with 81.5% of tweets labelled as such. This is unsurprising as many profane words were used as query terms to extract the tweets. *Exclusion* was found to be the least assigned label, accounting for only ten tweets. About a sixth of the tweets in the dataset belonged to the *None* class – i.e., tweets that were not assigned any labels and *bullying* was the fifth most prominent label. As described in Table 3.2, a tweet is only labelled *bullying* if, in addition to being offensive, it is targeted at an identifiable person or group; as such, not all tweets labelled as *profanity* or *insult* were labelled as *bullying*. For example, the tweet “*yall just nasty hos shut the f*ck up*” was assigned profanity and insult but not *bullying* as it appears to be directed at all women. In all, 82.8% of the tweets contained some form of offensive content which is higher than any existing cyberbullying dataset to date. The tweets count for each label is as shown in TABLE 3.3.

Label	Count
Profanity	51,1014
Porn	16,690
Insult	15,201
Spam	14,827
Bullying	3,254
Sarcasm	117
Threat	79
Exclusion	10
None	10,768

TABLE 3.3: Total number of tweets each label was assigned to.

Prior to preprocessing, the maximum document length for the dataset was 167 characters, with an average document length of 91 characters. After preprocessing, the maximum document length reduced to 143 characters (equating to 26 words), with an average document length of 67 characters. There are a total of 37,453 distinct word tokens in the dataset. Single label tweets make up more than a third of the dataset, and this can be attributed to the large number of tweets singly labelled as *Profanity*. Furthermore, a significant number of tweets were jointly labelled as *Profanity* and *Insult* or *Insult* and *Cyberbullying*, and this contributed to double-labelled tweets being the second-largest group of the dataset. Interestingly, there were more tweets associated with quadruple labels than there were with triple and this was discovered to be due to the

high positive correlation between *Porn + Spam* and *Profanity + Insult*. FIGURE 3.1 illustrates tweets counts and the number of tweets assigned.

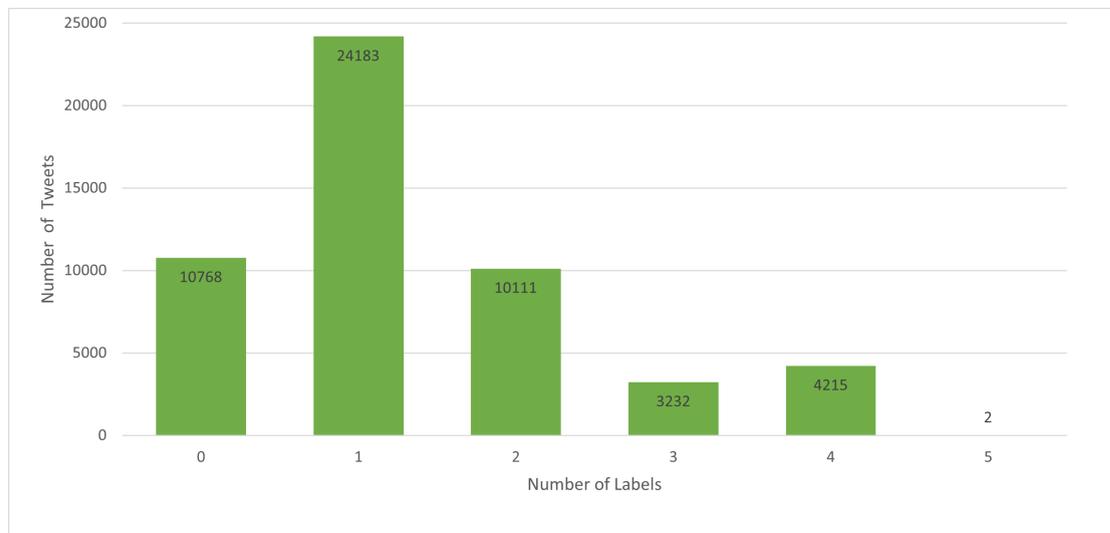


FIGURE 3.1: Distribution of tweet counts and number of labels assigned.

The correlation matrix for the labels is illustrated in FIGURE 3.2. *Porn* and *Spam* exhibited the highest positive correlation, and some positive correlation is also evident between the two labels and *Profanity*. Since pornography and spam share a lot in common and typically feature profane language, this correlation is not unexpected. Indeed, many pornographic tweets are essentially profanity-laden spam. *Insult* also exhibited a positive correlation with *Bullying* and *Profanity*, a fact that can be attributed to the frequent use of profanity in insulting tweets as well as the use of insults to perpetrate bullying. Only one negative correlation was identified in the dataset, and this is between *Bullying*, and *Porn + Spam*, implying a mutually exclusive relationship between the labels. Bullying tweets are often personal attacks directed at specific individuals and typified by the use of mentions (i.e., tagging another user in a tweet), person names or personal pronouns. By contrast, pornographic and spam tweets are devoid of these since they are rarely directed to specific individuals. This inverse relationship is evident in the dataset as no bullying tweet was classified as *Porn* or *Spam*. Minority classes like *Sarcasm*, *Threat* and *Exclusion* demonstrated no correlation with the other classes.

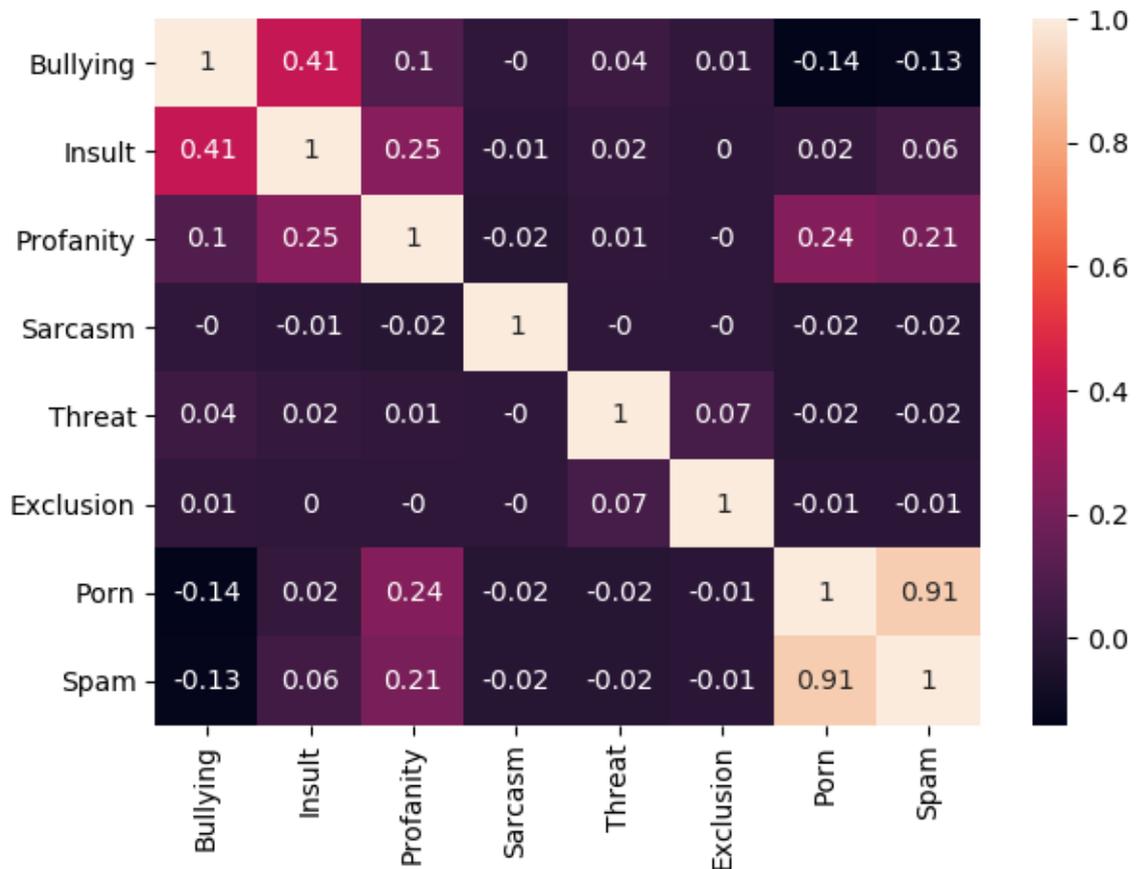


FIGURE 3.2: Distribution of tweet counts and number of labels assigned.

3.3.7 Bias and Practical Usage Implication

Most datasets carry a risk of demographic bias (Hovy and Spruit, 2016), and this risk can be higher for datasets created using manually-defined query terms (such as the one created as part of this research program). This bias can be in the form of gender, ethnic or cultural identities contained in the dataset, which may possess certain attributes that could be inadvertently adopted by the model and subsequently influence its predictions. For example, a National Institute of Science and Technology (NIST) study (Grother *et al.*, 2019) discovered that many US-developed facial recognition algorithms generated significantly higher false positives for Asian and African-American faces compared to Caucasian faces. Similar algorithms developed in Asian countries did not show any such dramatic differences in false positive rates between Asian, African-American and Caucasian faces. The study concluded that the use of diverse training data is critical to reducing bias in such AI-based applications. Researchers, therefore, need to be aware of

potential biases in datasets and address them where possible. Discovering potential biases within the created dataset (as far as possible) was therefore of interest as offensive content often have racial and sexist undertones.

Since Twitter does not collect users' gender information, secondary means were used to infer the gender of the users incorporated in the dataset. The Gender API⁹ was used to deduce users' gender based on whether the users' first names are traditionally male or female: which was assumed as an accessible and feasible measure of users' gender identity. The authorship of 13,641 tweets (21.8% of the dataset) could be processed in this way, and it was inferred that 31.4% of the authors of these tweets identified as female and 68.6% male (at least in so far as was apparent from their Twitter account).

This suggests a male-bias in the authorship of the tweets in the dataset. The limitation of this approach is acknowledged as the names provided by users cannot always be regarded as truthful, and as gender extends beyond the traditional binary types, a names-based approach such as this cannot be used to deduce all gender identities.

To mitigate potential racial and ethnic bias, variants of queries that used ethnicity-specific keywords were created for other ethnicities and used to retrieve tweets. For example, *#asianlivematters*, *#whitelivematters*, *#bluelivesmatter* and *#alllivematters* were all used as query terms to help mitigate any bias that may have been inadvertently introduced by the use of *#blacklivematters* as a query term. It should, however, be noted that the popularity and topicality of certain keywords may still introduce an unintended bias. For example, *#blacklivematters* returns several more tweets than *#asianlivematters*.

While the collection strategy used to create the dataset ensured a high concentration of offensive tweets, a potential consequence of the imbalanced distribution of the classes is that it may reinforce the unintentional bias of associating minority classes to specific hateful and abusive terms. Dixon *et al.* (2018) defined unintended bias as when a model performs better for comments containing specific terms over others. For example, the phrase "*stay in your lane*" was found in 4 of the 10 tweets identified as *exclusion* (due to the presence of the hashtag *stayinyourlane* in the tweet's content), and this can cause a model trained on the dataset to overgeneralised the phrase's association with the *exclusion* label, thus introducing a false positive bias in the model. To mitigate biases such as these,

⁹gender-api.com

the proposed cyberbullying prevention system allows users to reclassify messages and utilise the reclassified messages for online training of the ML models.

3.3.8 Dataset Availability

The final dataset of 62,587 tweets is publicly available as a csv file at <https://bitbucket.org/ssalawu/cyberbullying-twitter>. Twitter's usage terms restrict the mass distribution of tweet contents directly; as such, each tweet's unique ID is provided instead along with the assigned labels (see TABLE 3.4).

ID	Bullying	Insult	Profanity	Sarcasm	Threat	Exclusion	Porn	Spam
1134515	1	1	1	0	0	0	0	0

TABLE 3.4: Example row from the dataset.

3.4 Best Performing Model Selection

A number of traditional ML and deep-learning models were experimented with to perform multi-label classification on the dataset. The traditional ML algorithms used were Multinomial Naive Bayes, Linear SVC, and Logistic Regression while BERT, DistilBERT, RoBERTa, XLNET were the chosen deep-learning models. Standard performance metrics and observed predictions on unseen data were then used to select the best performing model to use in the cyberbullying prevention mobile application developed as part of this programme of research.

3.4.1 Models Evaluation and Best Performing Model Selection

Macro ROC-AUC (Area Under ROC Curve), Accuracy, Hamming Loss, Macro and Micro F1 Score, were the metrics selected to evaluate the models' performance. These metrics are typically used to evaluate a model's performance in imbalanced classification tasks and additionally, when computed alongside other performance metrics during several iterations of the experiments performed, these metrics returned consistent results.

Macro ROC-AUC (Area Under Receiver Operating Characteristic Curve) (Hanley and McNeil, 1982), Accuracy (Bratko, 1997), Hamming Loss (Destercke, 2014), Macro and

Micro F1 Score (Opitz and Burst, 2019) were the metrics selected to evaluate the models, as is standard for imbalanced classification tasks. A ROC curve is a graphical plot of the true positive rate against the false positive rate for a classifier's predictions at various threshold settings (see FIGURE 3.3). The area under this curve is the ROC-AUC value. When performing multi-label classification (as was the case in the experiments conducted), the macro ROC-AUC which combines the mean true positive and false positive rates for all the labels, is typically used.

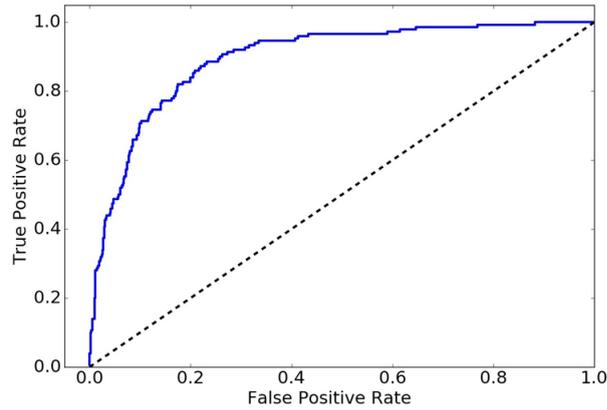


FIGURE 3.3: Example of ROC Curve.
Source: Choudhary *et al.* (2019)

Accuracy is defined as the proportion of correct predictions (both true positives and true negatives) among the total number of samples classified (Metz, 1978) and is represented by the formula:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

where TP = True positive; FP = False positive; TN = True negative; FN = False negative. Instead of computing the number of correctly predicted instances (like accuracy), the Hamming Loss reports on the loss generated in the predicting labels. It is the result of an XOR operation between bit strings representing the actual and predicted labels for a sample and is represented by the equation

$$HL = \frac{1}{NL} \sum_{i=1}^N \sum_{j=1}^L XOR(Y_{i,j}, \hat{Y}_{i,j})$$

where N and L represents the number of samples and labels respectively and $Y_{i,j}, \hat{Y}_{i,j}$ represents the actual and predicted bit representation of the label j in the data instance i . F_1 Score is the harmonic mean of precision and recall. The Macro F_1 Score (also referred to as the average F_1) is computed by calculating the F_1 Score for each class and is represented by the formula:

$$MacroF_1 = \frac{1}{n} \sum_x \frac{2P_x R_x}{P_x + R_x}$$

The Micro F_1 Score (also called the F_1 of averages) is calculated by computing the precision and recall for each class and calculating their harmonic mean (Opitz and Burst, 2019). It is represented by the formula:

$$MicroF_1 = 2 \frac{(\frac{1}{n} \sum_x P_x)(\frac{1}{n} \sum_x R_x)}{(\frac{1}{n} \sum_x P_x) + (\frac{1}{n} \sum_x R_x)}$$

where P_x and R_x are the *precision* and *recall* for each class in both equations. A classifier's precision is intuitively defined as its ability not to label a negative sample as positive while the recall provides a measure of its ability to find all the positive samples. Both metrics are represented by the formulas below:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

where TP, FP, and FN are as defined above for accuracy. Stratified K-fold cross-validation was used as the validation strategy. Cross-validation involves randomly shuffling the dataset into K groups and then training the model with K-1 groups and using the remaining group for testing. The process is repeated until each group has been used for testing, and the results averaged over all the runs. For an imbalanced dataset such as the one created, stratified K-fold is the preferred cross-validation method. This is because unlike K-fold cross-validation, which simply divides the dataset into K parts without consideration for the distribution of the dataset classes, stratified K-fold cross-validation

maintains the percentage distribution of classes across the K groups. Therefore, if K is 5 and class A contains 20 samples, the dataset is divided into 5 equal parts with each part containing 4 samples of Class A. As is standard for many NLP tasks, K was set to 10. This also ensures that one sample of the minority class (exclusion – which was only assigned to 10 tweets) is included in each generated subsets. In addition to the pre-trained models, fine-tuning was performed for each of the Transformer-based models using the text extracted from the tweets. The results of the experiments are presented in TABLE 3.5.

Model	Macro ROC-AUC (↑)	Accuracy (↑)	Hamming Loss (↓)	Macro F_1 (↑)	Micro F_1 (↑)
Multinomial Naive Bayes	0.8030	0.4568	0.1014	0.2618	0.7200
Linear SVC	0.8353	0.5702	0.0866	0.3811	0.7674
Logistic Regression	0.8354	0.5743	0.0836	0.3587	0.7725
BERT _{pre-trained}	0.9657	0.5817	0.0736	0.6318	0.7998
DistilBERT _{pre-trained}	0.9675	0.5802	0.0764	0.5202	0.7855
RoBERTa _{pre-trained}	0.9695	0.5785	0.0722	0.5437	0.8081
XLNet _{pre-trained}	0.9679	0.5806	0.0738	0.5441	0.8029
BERT _{fine-tuned}	0.9651	0.5822	0.0725	0.5300	0.8022
DistilBERT _{fine-tuned}	0.9633	0.5834	0.0753	0.5040	0.7872
RoBERTa _{fine-tuned}	0.9670	0.5794	0.0724	0.5329	0.8044
XLNet _{fine-tuned}	0.9654	0.5819	0.0741	0.5308	0.8037

TABLE 3.5: Results of classification experiments.
(↑: higher the better; ↓: lower the better)

The best macro ROC-AUC, Micro F_1 and Hamming Loss scores were achieved by RoBERTa_{Pre-trained}, while the best Macro F_1 and accuracy scores were attained using BERT_{Pre-trained} and DistilBERT_{fine-tuned} models, respectively. DistilBERT was the only fine-tuned model to achieve the best result for any metric. As expected, the deep learning models outperformed the baseline classifiers, with Multinomial Naive Bayes providing the worst results across the experiments. Interestingly, the pre-trained models performed better than the equivalent fine-tuned models, implying that fine-tuning the models on the dataset degrades rather than improves performance. This is similar to the results of the experiments conducted by Radiya-Dixit and Wang (2020), which discovered that fine-tuned networks do not deviate substantially from the pre-trained one. A possible reason for the performance degradation experienced when using the fine-tuned models could be due to the language gap between the datasets used for pre-training the models

(i.e., BooksCorpus (Zhu *et al.*, 2015) and Wikipedia) and the created Twitter dataset used here. Large pre-trained language models thus have high generalisation performance.

As would be expected, the models performed better at predicting majority classes (i.e., *Profanity*, *Spam*, *Porn*, *Insult*, *Bullying*) than minority classes. RoBERTa and XLNet performed better at predicting minority classes like *Sarcasm*, *Threat* and *Exclusion* than the other models. All the models performed well in predicting the none class, i.e. tweets with no applicable labels.

Overall, RoBERTa_{Pre-trained} emerged as the best performing model, achieving the best results in three out of the five evaluation metrics. Following the evaluation experiments conducted, all twelve classifiers were used to predict the 8 labels for a sample of 250 unseen offensive and inoffensive tweets and the results manually compared to identify from a "human perspective" which of the models provide predictions that are most acceptable for human consumption. A key consideration was how well the models discriminate between mildly offensive to very offensive tweets. RoBERTa_{Pre-trained} and XLNet_{Pre-trained} were the two models that excelled in this exercise. Consequently, RoBERTa_{Pre-trained} was the final model selected for use in the cyberbullying prevention mobile application since it performed as well as XLNet_{Pre-trained} in the "human perspective" evaluation and achieved the best results for three evaluation metrics while XLNet_{Pre-trained} in comparison did not achieve the best score for any of the evaluation metrics.

3.4.2 Evaluating the Dataset's Fitness for Purpose

As the proposed cyberbullying prevention system would be used to detect cyberbullying and online abuse on various social media platforms, it is essential that the knowledge acquired from the Twitter-based dataset can be used to detect cyberbullying on other SMPs like Facebook and Instagram. To assess the dataset's generalisability; cross-domain experiments were conducted to evaluate if the knowledge acquired by a model trained on the Twitter-based dataset can be used to perform similar tasks (e.g., cyberbullying detection) on another domain (e.g., Wikipedia). RoBERTa_{Pre-trained} as the best performing model from the earlier experiments was used to predict the labels on two other unseen datasets. New instances of RoBERTa models were then trained on

these two datasets and used to predict labels for the Twitter-based dataset and the results of the two set of experiments compared.

The dataset created by Davidson *et al.* (2017) and the Kaggle Toxic Comments dataset Kaggle (2018) were selected for the experiments; herein referred to as the Davidson (D) and the Kaggle (K) datasets, with the newly created dataset referred to as the Cyberbullying (C) dataset. The Davidson dataset is a multi-class-labelled dataset sourced from Twitter where each tweet is labelled as one of hate_speech, offensive and neither. In contrast, the Kaggle datasets contained Wikipedia documents labelled using a multi-label annotation scheme with each document associated with any number of classes from toxic, severe_toxic, obscene, threat, insult, identity_hate. Due to the difference in the number of labels for each dataset (Cyberbullying dataset – 8 labels, Davidson dataset – 3 labels, Kaggle dataset – 6 labels), it was necessary to amend the input and output layers of the RoBERTa model to allow it to predict the relevant labels for the Davidson and Kaggle datasets.

For the reverse experiments, new instances of RoBERTa were trained on both the Davidson and Kaggle datasets and used to predict labels for the Cyberbullying dataset. As control experiments, in-domain evaluation was also performed. This involved training RoBERTa models on the Davidson and Kaggle datasets and using these to predict the labels on the same datasets. The results of the experiments are presented in Table 3.6.

Model	Macro ROC-AUC (↑)	Accuracy (↑)	Hamming Loss (↓)	Macro F_1 (↑)	Micro F_1 (↑)
RoBERTa _{C→D}	0.9923	0.8809	0.0288	0.8802	0.8810
RoBERTa _{D→C}	0.9681	0.5831	0.0708	0.5330	0.8076
RoBERTa _{D→D}	0.9905	0.8814	0.0300	0.8427	0.8758
RoBERTa _{C→K}	0.9916	0.5924	0.0123	0.5670	0.7436
RoBERTa _{K→C}	0.9651	0.5811	0.0727	0.5352	0.8054
RoBERTa _{K→K}	0.9733	0.8449	0.0174	0.5026	0.6354

TABLE 3.6: Results of cross-domain experiments.
(↑: higher the better; ↓: lower the better)

Overall, models trained on the Cyberbullying dataset (RoBERTa_{C→D} and RoBERTa_{C→K}) perform better on the Davidson and Kaggle datasets than the models trained on these two other datasets and tested on the Cyberbullying dataset (RoBERTa_{D→C}, RoBERTa_{K→C}). Interestingly, models trained on the Cyberbullying dataset achieved better ROC-AUC, Macro and Micro F_1 values on both the Davidson (D) and the Kaggle (K)

datasets compared to in-domain results on those datasets (i.e., models trained and evaluated on the same datasets – $\text{RoBERTa}_{D \rightarrow D}$ and $\text{RoBERTa}_{K \rightarrow K}$). The results indicate that the Cyberbullying dataset sufficiently captures enough context for classifiers to distinguish between cyberbullying and non-cyberbullying text across different social media platforms. Even more impressive is that the cross-domain results achieved by the RoBERTa model trained on the Cyberbullying dataset, which comprises short-form content from Twitter, are better than the in-domain results achieved by another RoBERTa model trained on the same long-form based dataset (the Davidson dataset).

3.5 Summary

The standard practice when creating cyberbullying datasets is to emulate the real-world distribution of cyberbullying on social media, with researchers going to great lengths to emulate this distribution as closely as possible. The research presented here proposes a novel approach that differs from standard practice. Using a series of targeted queries executed on the Twitter Search and Streaming APIs, potentially abusive and bullying tweets were extracted and annotated. The resulting dataset contained over 82% abusive and offensive content, and two sets of experiments were performed to determine the dataset's suitability for training classifiers. The first group of experiments evaluated a variety of traditional and deep learning classifiers and was used to identify the best performing model to use in the cyberbullying prevention mobile app developed as part of this research. This was found to be the pre-trained RoBERTa model.

The second set of experiments was aimed at determining the new dataset's suitability for training models to predict online abuse on different social media and messaging platforms. The results are extremely positive and validate the dataset's suitability for training models to generically detect online abuse. The research presented in this chapter is novel and represents the first attempt to create a cyberbullying and offensive language dataset with such a high concentration of cyberbullying and offensive language. Furthermore, a model trained using the dataset achieved better performance on a different domain than the in-domain performance of a model trained on the target domain.

The implication is that the dataset's imbalanced nature did not affect the model's ability to learn both offensive and non-offensive content. Instead, it improved the model's

performance and, in some instances, surpassed in-domain results on different datasets. Moreover, the effort that would have been expended in simulating a real-word distribution for the dataset has been saved and oversampling to improve the proportion of bullying content within the dataset is not required given the high proportion intrinsic to the dataset.

The recent success and popularity of deep learning models in many areas of machine learning have designated them state-of-the-art for natural language processing tasks. The findings from the experiments conducted as part of this reported research reaffirm this assertion. The deep learning models outperformed traditional classifiers and RoBERTa – an optimised BERT-like model – emerged as the best performing classifier. This is in agreement with experiments performed in other studies where RoBERTa outperformed other Transformer-based models like BERT, DistilBERT and XLNET; it, therefore, influenced the decision to use it as the classifier in the novel cyberbullying detection system developed as part of this programme of research.

Chapter 4: Stakeholders' Perspectives on Cyberbullying Prevention

4.1 Introduction

As discussed in Chapter 2, there have been considerable research efforts devoted to the development of machine learning algorithms to detect cyberbullying and online abuse; unfortunately, the scientific advances represented by these algorithms are yet to result in a proliferation of practical software applications for cyberbullying mitigation and prevention. Existing cyberbullying prevention systems such as those developed by Lempa *et al.* (2015), Weider *et al.* (2016), Vishwamitra *et al.* (2017), Talukder and Carbanar (2018) and Shome *et al.* (2019) (see Section 2.7.3) suffer common shortcomings such as poorly implemented integration to online social networks, an inability to understand language contextually (due to a reliance on wordlists to detect offensive language) and usability concerns. Some of these failings are due to a poor understanding of end-user requirements and a lack of consideration for users' opinions during the design process.

The literature on technology-based prevention of cyberbullying is conspicuously devoid of discussion on how systems' requirements were captured and the means by which potential users were engaged (if at all) in the design and evaluation process. This highlights a failing in existing attempts to develop cyberbullying and online abuse prevention tools that are fit for purpose, and that effectively meet end-user needs. Capturing the desires, needs and aspirations of end-users should be an integral part of the design process for any interactive tool – as noted by Maguire and Bevan (2002, pg. 133) “*successful systems and products begin with an understanding of the needs and requirements of the users*”. Adopting a User-Centred Design (UCD) approach is crucial to achieving this understanding. UCD is

a system design approach that emphasises the importance of the end-user to the design process and identifies that the purpose of a system is to serve the user. Norman (1986, pg. 61) succinctly summarised the relationship between end-users and a system when he observed that “*the needs of the users should dominate the design of the interface, and the needs of the interface should dominate the design of the rest of the system.*” The first consideration when applying UCD methods is to therefore understand the end-users and the potential interactions that will occur between them and the system.

This chapter reports on the activities undertaken as part of this research to gain an understanding of users’ requirements for the proposed mobile application. As identified in Chapter 1, the key stakeholders for the mobile app include parents, educators, law enforcement, mental health professionals and adolescents; collectively, these are broadly categorised as adult and young stakeholders. Due to the difference in age and the sensitive nature of the topics being discussed, different UCD methods were used with the two groups. Questionnaires and interviews were the tools used to engage with the young participants to gain insight into their views and experiences of cyberbullying. For the adult stakeholders, a number of focus group sessions were held to discuss pertinent issues about cyberbullying and how it can be prevented using a tool like the proposed mobile app. The rest of the chapter is divided into three main sections; Section 4.2 presents the discussions and findings from the focus groups, and Section 4.3 explores adolescents’ views on cyberbullying and its prevention. Finally, Section 4.4 provides a summary of the chapter’s discussions

4.2 Adult Stakeholders’ Perspectives on Cyberbullying Prevention

While adolescents are the primary target audience for the mobile app, adults entrusted with their care, mental wellbeing and protection are other stakeholders. Such adults include parents or guardians, educators, law enforcement and mental health professionals. The views of parents and guardians of young children are of interest to the research programme due to their influence on the young people who are the intended audience of the proposed cyberbullying prevention application. They are also likely to

procure mobile devices for young people and are thus well placed to recommend a cyberbullying prevention tool like the one proposed to young people.

Teachers are part of schools' prevention strategies for cyberbullying and are often entrusted with implementing prevention guidelines. They may have also witnessed cyberbullying incidents and thus provide a different perspective from parents and guardians. Severe cases of all forms of bullying (including cyberbullying) and its negative consequences are often referred to mental health specialists. Clinicians could therefore contribute their knowledge on cyberbullying involvement for young people, how reported cases are managed from a clinical perspective and help explore ways by which some of the clinical strategies can be implemented in the proposed mobile application. Finally, as cyberbullying is considered an electronic misdemeanour, law enforcement officers, particularly cybercrime specialists, can provide a unique perspective borne out of their experience enforcing the applicable laws for cyberbullying.

Incorporating the views of this group of stakeholders into the design process and winning their trust in the process will hopefully lessen the adoption barrier for the mobile app. The following subsections discuss the design, implementation, findings and limitations of a focus groups study conducted to understand adult stakeholders' attitudes to cyberbullying and its prevention using technology.

4.2.1 Focus Group and Participant Recruitment

A focus group study is an organised discussion with a group of individuals to gain information about their views and experiences in relation to the topic(s) of discussion. Focus groups are used in UCD approaches to gather end-user requirements and gain a deeper understanding of the problem being studied. The merits of focus groups as a means of understanding users' requirements for a system have been espoused by studies such as those of Biediger-Friedman *et al.* (2018), Tresser (2017) and Harder *et al.* (2017). Biediger-Friedman *et al.* (2018) used focus groups to understand participants' current use of healthy living apps as the initial phase of their UCD approach to designing an app to promote healthy lifestyles; they discovered that participants' willingness to use health apps is dependent on the app's ease of navigation and the inclusion of specific features like health tips targeted at their specific demography (e.g., English- and

Spanish-speaking nursing mothers). They were then able to leverage this information in the later phases of the app development. Similarly, Tresser (2017) used interviews and focus groups to engage with clinicians and technology experts to gather requirements for a Virtual Reality (VR) game designed to assist in the physical therapy of children with cerebral palsy. Through the approach, they discovered the importance of practising sustained upper body and trunk movements and consequently concentrated their development efforts on these areas in their VR game.

In Harder *et al.* (2017), users' requirements for a mobile app to support post-surgery recovery in breast cancer patients were gathered using focus groups. The researchers were able to capture a prioritised list of application features desired by the users, which was used to improve the user experience. Within cyberbullying literature, focus groups have been used to understand parents' and teachers' perspectives on cyberbullying (Monks *et al.*, 2016; Jäger *et al.*, 2010) as well as its impact on young people (Agatston *et al.*, 2012; Smith *et al.*, 2008). Gibbs (1997) posited that, by providing participants with an opportunity to discuss research topics with each other and researchers, a focus group could be an empowering experience for participants. It can imbibe a sense of pride for being valued as experts and contributing meaningfully to the research (Goss and Leinbach, 1996), which can make participants more inclined to share personal views with other individuals. Focus groups were, therefore identified as the ideal vehicle to engage the adult stakeholders to elicit their various perspectives on cyberbullying and how the proposed mobile app can assist in its prevention.

Before the commencement of the study, ethical approval was sought and granted by the Aston University Research Ethics Committee for all stages of the study (see Appendix B.1). There were four types of adult stakeholders recruited as participants: mental health professionals; educators; law enforcement officers; and parents. To recruit the mental health professionals, the researcher initiated contact with local Child and Adolescent Mental Health Services (CAMHS) units. CAMHS is an NHS service that assesses and treats young people with emotional, behavioural or mental health difficulties including depression, bipolar disorder, schizophrenia, self-harm, anxiety, abuse and many more. This engagement was facilitated by a personal contact who worked in one of the units. Five clinicians indicated an interest in taking part in the study, and each was sent an invitation email (see Appendix B.2) and a Participant Information Sheet (PIS) (see Appendix B.3) that provided information about the focus group study and the overall

research objectives. Three clinicians (a psychiatrist, a child psychologist and a mental health nurse) confirmed that they would like to participate in the study. The researcher met with the clinicians to provide more information about the research and to understand how the services offered by CAMHS are provided to the community, especially with regards to cyberbullying concerns raised by young people and their parents. These discussions revealed the following information:

1. the primary means to access CAMHS services is via a referral from the patient's local general practitioner (GP);
2. some CAMHS units will also accept referrals from schools and social workers; and
3. some GPs with mental health specialisms can provide initial care for bullying-related complaints.

The above points highlight the crucial roles played by local GPs and educators in providing young people dealing with cyberbullying (as victims and offenders) with access to professional mental health services. While teachers have already been identified as stakeholders for the cyberbullying prevention mobile application, GPs were not initially included. Based on the discovered information, it was decided that the inclusion of a GP as a participant would be beneficial to the study. Personal contacts were then explored to this end, and a GP with mental health specialism was successfully recruited to take part in the study.

Running in parallel to the clinicians' recruitment were efforts to engage local secondary schools to recruit teachers as participants. Secondary schools in Birmingham, Wolverhampton and Walsall, were contacted to participate in the study. In addition to this, the researcher reached out to personal contacts in the teaching profession to publicise the study as part of efforts to engage educators. These activities resulted in two schools agreeing to take part in the research and four teachers and a private tutor getting in touch with the researcher to indicate their interests.

With regards to the engagement with law enforcement, the researcher contacted the West Midlands Police Research Unit, and an official request for research support was submitted (see Appendix B.4). The proposal was subsequently approved, and the Cyber Crime Manager for the West Midlands Regional Cyber Crime Unit (a Detective Inspector) assisted with the study.

The final type of stakeholders were parents. The researcher attended a number of school assemblies to inform the students about the research and seek their assistance in recruiting parents for the focus group study. The PIS was given to the students to take home to their parents, but after several unsuccessful attempts to recruit parents through the students in this manner, the decision was made to use a convenience sampling method for the parents. The lack of interest amongst the schools' parents was believed to be due to the sensitive nature of cyberbullying and that many parents may not be comfortable discussing such a topic with a stranger (i.e., the researcher). By utilising personal contacts, this barrier would be significantly reduced. Moreover, the parents' within the convenience sample are essentially the same demography as those being recruited through the schools (i.e., parents with children in local secondary schools). The study was therefore publicised (via WhatsApp) to parent groups that the researcher belonged to (as a parent) such as school and extra-curricular activities groups and a total of eighty-two parents were made aware of the research (based on the membership of these groups). Nine parents indicated an interest in the study and were sent invitation emails and the PIS. The researcher then spoke with the nine parents over the phone to answer questions and provide additional details about the study.

While research has been conducted to understand adults' perception of cyberbullying (Dehue *et al.*, 2008; Eden *et al.*, 2013; Makri-Botsari and Karagianni, 2014), rarely has all the stakeholder groups (i.e., parents, teachers, clinicians and law enforcement) identified by this study been engaged by a single study to explore their opinions together within the same gathering. The study by Moreno *et al.* (2018) was one of the few discovered that attempted this, and their engagement was focused on exploring the stakeholders' understanding of cyberbullying as a means to validate the uniform definition of bullying proposed by Gladden *et al.* (2014). This research is, therefore, novel in its engagement with the stakeholders to devise strategies to prevent and mitigate cyberbullying through the use of an automated online abuse prevention tool.

All prospective participants were contacted to agree on a schedule and dates for the focus group sessions. Unfortunately, some prospective participants (including the two secondary schools) withdrew from the study due to scheduling conflicts. Overall, eleven participants comprising three clinicians (a psychiatrist, a child psychologist and a GP), three educators (2 teachers and a private tutor), four parents and a law enforcement officer took part in the study. All participants were, coincidentally, parents, but only seven had a child aged 11

years or older. While a focus group size of 10-12 participants is often advocated (Stewart and Shamdasani, 2014), studies such as those by Kitzinger (1995), Krueger (2014) and Halcomb *et al.* (2007) proposed smaller group sizes (4 - 8), especially when working with a single moderator. Likewise, Barbour and Kitzinger (1998) believed that a focus group of more than six participants might be too large for sociological studies and that researchers should instead be flexible with regards group sizing as factors outside of their control may limit the feasible size of the group. This recommendation was adopted by the study and ten participants (not including the law enforcement offer) comprising the seven female, and three male participants were divided into two groups of five (designated Groups A and B), and each group contained both male and female participants.

Although three sessions were planned for each group, due to the limited availability of some participants, Group B only met twice. The objectives of the planned second and third sessions for Group B were thus covered during the second meeting, the duration of which was extended to accommodate this. A total of five sessions were therefore held across the two groups. Due to the restricted availability of the law enforcement representative, the officer worked with both focus groups, attending two sessions with Group A and the final (extended) session with Group B. The sessions were held between October 2017 - February 2018.

In assigning participants to the groups, it was ensured that key stakeholders types were represented within each group. Each group, therefore, included at least one of the following:

- Parent with a child of 11 years or older;
- An educator; and
- A mental health professional.

All participants signed consent forms (see Appendix B.5) and, although they were offered a £10 Amazon voucher per session (as indicated in the PIS) in token appreciation of their time, all participants waived this.

Each session lasted between one and a half to two hours and was audio-recorded and transcribed. The researcher provided hot beverages and snacks for all sessions. Content analysis was performed based on the work of Marshall and Rossman (2014), Moretti

et al. (2011) and Wilkinson (2011) and is discussed further in Section 4.2.2. Each session started with a review of the study's objectives, followed by a discussion of the session's specific objectives (see Appendix B.6) to help focus participants. Additionally, the researcher/moderator thanked participants for their time and reinforced to them the importance of their contributions in furthering cyberbullying prevention research.

As the first phase of a user-centred design approach to the development of a mobile-based cyberbullying prevention tool, the focus group sessions were aimed at gaining a good understanding of the adults' perception of cyberbullying, its perpetuation on social media and their views on how it can be prevented. Another objective of the focus groups was to understand how much control parents exert over their children's use of mobile devices, and if they feel this is an effective strategy for mitigating cyberbullying. The first sessions for both groups were focused on gaining an insight into participants' perceived significance of cyberbullying and what they understood to be its critical issues. The second sessions explored cyberbullying prevention strategies and participants' experience dealing with cyberbullying and implementing prevention strategies, while the final session of Group A and second half of Group B's final session focused on the proposed cyberbullying prevention mobile app and the participants' desired features for the app.

4.2.2 Content Analysis

Before coding could commence, a unit of analysis had to be identified. A unit of analysis is a comprehensible piece of a transcript that contains an idea or a piece of information (Schilling, 2006) and is used as the basis of developing a coding system (Wilkinson, 2011). Marshall and Rossman (2014) proposed a six phased thematic analysis process, as shown in TABLE 4.1, which maps these six phases to the actual tasks performed by the researcher at each stage of the process. NVivo software was used to store, manage and facilitate the analysis of the qualitative data.

4.2.3 Emergent Themes

Content analysis of the focus group data revealed four major themes. These are discussed in detail in the following sections and the coding tables are presented in Appendix B.7.

	Phase	Tasks
1	Organise data	<ul style="list-style-type: none"> • Transcripts across both groups were collated in chronological order – i.e., Session 1 transcripts for both groups were placed together. • Transcripts were read carefully several times to gain a clear and accurate sense of the discussions.
2	Generate categories or themes	<ul style="list-style-type: none"> • Highlight phrases that contain an idea or important information about the topic and assign a code.
3	Code the data	<ul style="list-style-type: none"> • Group highlighted phrases that expressed similar concepts into categories. • Review the categories, merge related categories and formulate new ones if required. • Organise linked categories into a hierarchical structure and review again to identify overlaps or if further division is necessary.
4	Test emergent understanding of the data	<ul style="list-style-type: none"> • Place emerging themes within the context of existing theories from the literature or create new theories.
5	Search for alternative explanations of the data	<ul style="list-style-type: none"> • Challenge current understanding of the data and explore alternative explanations
6	Write up the data analysis	<ul style="list-style-type: none"> • Create a report of the analysis using extracts from the data as well as the literature to support interpretations.

TABLE 4.1: Thematic Analysis Phases and Tasks.

4.2.3.1 Concerns About Cyberbullying

The participants demonstrated a high level of awareness of young people's engagement with social media and how this can expose them to cyberbullying and other forms of online abuse. From the discussions in the first sessions of both groups, it was clear that many parents regarded cyberbullying as a major concern in relation to their children's social media usage. Some of the parents in the groups said:

"It seems like it's just everywhere now, especially on Twitter and Facebook. I haven't experienced any myself or my kids, at least that I know of but I see a lot of nasty comments all over the place even on LinkedIn which is meant to be like a sensible and professional site".

"There is more about it on the news now, which I think is good. I saw on the news about a girl that was being bullied by her mates, they were sending nude pictures of her, but the funny thing is they weren't even her pictures. They just got some porn pic and cut her face on it, just like that. I felt that was just mean, and these are like 12, 13 year olds".

I get so worried and anxious with all the stuff online, sometimes I see things on some pages and [I'm] like wow".

This high level of awareness exhibited by participants is a welcome improvement over findings of earlier studies such as that of Bauman (2010) and is in concert with the discoveries of recent studies such as those of Macaulay *et al.* (2020) and van Verseveld *et al.* (2020) which found teachers to be aware of the dangers posed to young by cyberbullying people. This is indicative of the changing awareness and attitudes of adults, particularly parents and teachers, with regards to young people's involvement with cyberbullying. Two participants reported that their children had been victims of cyberbullying and that they got involved in resolving the situation, with one parent noting:

"She didn't tell me about it for some time, and I had to prod her a bit, and it turns out it's even one of her close friends that I'm friendly with the mom, so I got the mum to have a word [...] I don't think they are friends anymore, but she stopped sending silly stuff to my girl anyway, and that's all I cared about, to be honest".

While none of the parents felt their children could have been involved in cyberbullying as the perpetrator, one participant admitted discovering their child sharing an inappropriate Internet meme about another child in school, noting:

“He thought it was just funny and not a big deal, but we talked, and I got him to think about how he would feel if he was on the receiving end of something like that. I think there’s an element of them not fully understanding the impact of some of their actions on the Internet”.

This inability to fully comprehend the negative consequences of sharing inappropriate material on the Internet was identified by Smith *et al.* (2008) as one of the factors contributing to young people’s involvement as perpetrators of cyberbullying. While the number of cyberbullying incidents reported by participants was admittedly low, there was no apparent gender-related relationship in the children’s involvement as either victims or perpetrators. This is in line with the findings of Hinduja and Patchin (2008) and Tokunaga (2010) who discovered no relationship between gender and the likelihood of being a cyberbully or victim but is in contrast to the conclusions of others such as Foody *et al.* (2019), Griezel *et al.* (2012), Baldry *et al.* (2016) and Navarro and Jasinski (2013) who uncovered gender-related relationships. Some parents admitted to being anxious over the dangers posed by cyberbullying to their children. Two parents said:

“I get so worried and anxious about how I can protect my kids from it. I’m like am I doing enough because sometimes you see things on some pages and [I’m] like wow, why would you put up something like this”.

“Once you’re being bullied on Facebook, maybe you should go off from Facebook, and then it can come from Twitter. It can get [to be] too much. It’s everywhere”.

A similar view was shared by another participant who advised his children to adjust their behaviour online to reduce the chance of them becoming cybervictims, saying:

“I tell them not to post pictures of themselves or things that can make others jealous or get involved in debates online”.

In discussing the pervasiveness of cyberbullying, a participant that works as a psychiatrist said:

“It is surprising how many cases of cyberbullying I see in the clinic. To be honest, when I started, I didn’t expect I will see this many, and some are quite severe”.

Just as reported in Bauman (2010), cyberbullying incidents often go unreported in schools. The teachers in the group confirmed as much, with one teacher noting:

“I’ve probably seen more fights and maybe just normal bullying. The students, they don’t want to be a snitch, so they don’t really tell teachers. And that’s just how it is sometimes, actually in my school, we have had quite a lot of assemblies and seminars about sexting, grooming and cyberbullying. The thing is we like to think we don’t have a bullying problem, but if I’m honest I don’t know”.

Whilst another added:

“We have had fights that started because of something posted online. They do that a lot, that’s actually quite common, funny enough. So you could say we end up with the aftermath of all the cyberbullying in the school. We sometimes call the parents in, but that doesn’t mean it ends there”.

4.2.3.2 Current Strategies and Solutions are in Need of Improvement

In exploring participants’ views on preventing cyberbullying and the effectiveness of the common strategies in use, it was evident that participants felt that social media companies are not doing enough to protect their young users from online abuse. Similar findings were reported in a study conducted by the charity organisation – The Children Society – which found that 83% of its 1,089 respondents wanted social media companies to do more to tackle the problem (BBC News, 2018).

Similar sentiments were expressed by many of the parents with regards schools’ efforts on cyberbullying prevention. A participant recounted a cyberbullying incident in which her daughter was a victim and said that she had *“to make a big deal”* to compel the school to act. The clinicians, however, sympathised with school authorities and highlighted the complicated nature of cyberbullying intervention as illustrated by the following quotes:

“It’s quite difficult for the schools sometimes because it’s constantly changing. The only thing they can do is to refer to us and then when you try to engage the parents, things don’t always go the way you expect them to, and you have to be very sensitive”.

“The schools have to stay objective as no parents will readily agree that their child is bullying another. No one wants to be the bully’s dad or mom, there is a stigma to that as well so, you know. It’s a bit complicated”.

“The children are aware that because the schools have no control over the Internet, they can bully more people over social media than if they have to walk up to them in school”.

From a law enforcement perspective, the cybercrime officer commented that the number of cyberbullying complaints received is significantly lower compared to other forms of cybercrime like online fraud, sexual grooming, and revenge porn. He added that complaints are usually made by parents or guardians but are often withdrawn after some time as the threat of prosecution serves as a good deterrent and typically dissuades the perpetrators. He spoke of the general advice provided to parents in such situations:

"We tell them to take screenshots, and as a first step they can send it back to the bully to say I have taken pictures of your texts and sent them to the police and nine times out of ten, that puts a stop to it".

The use of software as an online safeguarding strategy was uncommon amongst participants with only two participants reporting any experience of this. They said:

"I installed this app on [child]'s phone, it was meant to send me alerts when he uploads pictures on Instagram and things like when he sends messages and his location, but it kept crashing the phone, so we took it off. I'm sure he was very happy".

"I used the Vodafone parental control on our broadband, just because I felt I had to do something. Like I can't just let them be browsing without any form of control. It blocks site and stuff".

The remaining participants admitted to not having thought of using software applications in this manner:

"Nothing really, I wasn't aware there were things like that, to be honest".

"I didn't even think there would be an app to do something like that".

"I use to check their phones, but I never found anything. I would have probably used one of these [software] if I knew about them".

While these responses suggest low awareness of the use of technology for the protection of children online, it is worth noting that parents will generally not consider the use of such tools unless cyberbullying and risky Internet use is a concern. Overall, while participants showed a high level of awareness of cyberbullying and its prevalence on social media, many are unsure of what steps to take to proactively protect young people from its many

dangers. The general strategy for tackling cyberbullying appears to be blocking and ignoring cyberbullies, and in more severe cases, school authorities, law enforcement and mental health professionals tend to get involved. Finally, while participants concur that advice on how to manage cyberbullying situations is available online, they would like such advice to be presented in a more “*easily digestible*” form like a “*cheat sheet*”.

4.2.3.3 Encouraging Positive Behaviours and Online Safeguarding are Key Features for the Proposed App

All participants were enthusiastic about the prospect of an app designed to help young people combat cyberbullying and emphasised the importance of features that protect children when they are online and those that promote positive online behaviours. The clinicians particularly welcomed the idea and volunteered their time to assist further in the development of the proposed mobile app. The researcher presented an overview of the proposed app, as well as screenshots of some parental monitoring and social media apps to serve as visual cues for the participants. Discussions were focused on the functionalities that participants would like included in the app. The features suggested by participants can be broadly grouped into two categories, namely features that allow reflection, educate and empower users (discussed in this section), and features that apply punitive actions against abusive users (discussed in the next section). The following quotes illustrate the participants’ suggestions for features in the reflective, educational and empowerment category:

“I think like a safe browsing option will be good, so anything offensive is not shown to you when you are on Facebook and the likes”.

“I once read about having like a time out period from mobile phones and social media [...] I like that idea if you can put that in the app, it can just block out social media for like an hour or something”.

“[...] maybe it can rate children on how well they behave online”.

“It will be good if you can add links to some educational stuff about cyberbullying. It would be nice having all the information in one place”.

“I have an inspirational quotes app that I read in the morning. If the app can show something like that every day”.

“It can include some videos on how to treat people when they are online ”.

As all participants were parents, it was unsurprising that their suggestions centred around protective and reflective features. An opinion shared amongst many of the participants was that children often engage in online abuse unintentionally and under peer pressure. They, therefore, felt that tools that can highlight inappropriate behaviour would be beneficial. Additionally, participants favoured features that can protect young people while on social media, such as hiding inappropriate content from their news feed.

4.2.3.4 Report and Block Online Abusers

While participants favoured features aimed at encouraging positive attitudes amongst young people, they also suggested a number of the punitive actions that the mobile app could take against cyberbullies, including:

“[...] report people to Facebook so they can be banned [...]”.

“If it can automatically block Internet trolls, I think that would be great”.

“ [...] for really serious cases maybe report to the police or even just send them a text that you will be reported to the police”.

Interestingly, only one participant would like the app to include the ability to monitor a child's phone remotely. This suggestion was, however, opposed by other participants who felt the use of content filters to restrict the type of websites that can be accessed on the phone would be more appropriate.

4.2.4 Discussion

The focus group study was the first phase of a user-centred design approach to developing a cyberbullying prevention mobile app using participatory design. It was aimed at gaining an insight into adult stakeholders' understanding of cyberbullying, its prevalence on social media and how the proposed mobile app might assist in its mitigation and prevention. Throughout the study, the participants demonstrated continuous improvement in terms of their knowledge and understanding of cyberbullying, both via interaction with other participants and personal research. One participant declared:

"I feel like I'm an expert on cyberbullying now. There's so much going on that I didn't know about".

Another said that she has become more observant about her children's online habits, and she attributed this to the focus group discussions. The law enforcement officer concurred and said that the sessions have made him more "*sensitive to what young people go through*". He noted that:

"When we think of cybercrime, on the force, we mostly think in terms of financial costs and the links to organised crime, but really this stuff needs to be put at a similar level".

This was an unplanned but very positive corollary of the focus groups, the immediate impact of which was an improved level of engagement and contributions from all participants as the sessions progressed. The hoped-for long term impact is the cascading effect that these participants could have on other adults in the community with regards attitudes to cyberbullying and its prevention, and that could be a significant contribution to the fight against cyberbullying.

The study uncovered some interesting findings with regards adults' attitude to cyberbullying. Adults, particularly parents, displayed a good awareness of cyberbullying and the risks young people could be exposed to on social media. This awareness is, however, not matched by knowledge on how to proactively protect young people against online abuse or how to manage cyberbullying situations when they occur. While participants acknowledged that there is information available about cyberbullying on the Internet, many indicated a preference to have this information presented in a simple, digestible format like a cheat sheet. In subsequent interactions after the study, the researcher provided some online resources such as the downloadable PDF guides available on the BullyingUK website¹, and the participants responded positively to the guides' format. The implication of this is that advice on preventing and managing cyberbullying and online abuse is best presented in a question and answer format similar to a FAQ. This can assist anti-bullying organisations' efforts in raising awareness by making information resources more accessible.

The majority of the participants had not previously used software programs to manage risks associated with the Internet and social media, and many were not aware that it is

¹bullying.co.uk/

possible to use a computer program for such purposes. There are two key implications from this finding. Firstly, there is a lack of software tools to assist in cyberbullying prevention and, secondly, there is minimal awareness of the ones that do exist. Efforts to tackle this paucity by developing more automated tools should also, therefore, be accompanied by campaigns to make the public aware of their existence.

The participants reacted positively to the proposed mobile app, and the clinicians believed that the proposed app could be very useful in their work as a tool to assist cyberbullying victims to use social media safely. It is interesting that participants' initial suggestions for application features are reflective, and punitive elements were only suggested afterwards. The implication is that the preferred cause of action for adult stakeholders is to encourage positive attitudes amongst young people first and to only apply sanctions if this was unsuccessful. This helped form a conceptualisation of the kind of features desirable in the proposed app from the adult stakeholders' perspective, as outlined in TABLE 4.2 below.

Reflective, educational and empowerment	Punitive	Others
<ul style="list-style-type: none"> • Safe browsing mode when using social media apps • Provide links to educational material including videos • Display daily motivational quotes • Social media “time out” • Online behaviour scorecard 	<ul style="list-style-type: none"> • Block offensive users • Report offensive users to the social network • Report offensive users to law enforcement • Threaten offensive users with legal prosecution • Content filters to restrict access to inappropriate websites 	<ul style="list-style-type: none"> • Work with multiple social media platforms

TABLE 4.2: Adult stakeholder's desired features for the proposed app.

4.2.5 Study Limitations

The focus groups successfully contributed to the knowledge about adults' attitudes to cyberbullying and its prevention and the capabilities required in the proposed cyberbullying prevention application to gain their acceptance. While attempts were made to engage with the broader population of parents and teachers, these were ultimately unsuccessful, and personal contacts were used to facilitate recruitment. As the study utilised a convenience sample, the implication is that the participants' views may not be entirely representative of the wider population of parents, teachers, clinicians and law enforcement officers.

The participants' share similar views on cyberbullying and its prevention. They perceive cyberbullying as a concern and an online risk for young people, and as such, they are in favour of its prevention. This would have contributed to their overwhelmingly positive response to the proposed app. Adults with different views on cyberbullying and its prevention may not share a similarly positive attitude towards the proposed app. Their expectations and requirements for the app may therefore be different, and as such, there is a limitation that the views captured by the study are only reflective of adults positively disposed to preventing cyberbullying. Furthermore, the researcher's presence as the moderator in all sessions may have also encouraged an overall optimistic outlook about the proposed app's prospects from participants.

That said, ensuring a relaxed and positive atmosphere is critical to the success of a focus group (McLafferty, 2004), and the researcher's presence achieved this, allowing for an engaging discussion amongst participants. Cyberbullying is a sensitive subject, and as discovered by the researcher in the recruitment efforts, one that not many people are willing to discuss openly with strangers. A study of this nature requires well-motivated participants to openly discuss their views and concerns about the topic and commit to attending the multiple sessions required. It is, therefore, doubtful that these type of participants could have been recruited without shared trust and some familiarity.

A crucial part of enabling this is creating the right environment, and from the output of the study, the researcher was successful in making this possible. The focus group study was aimed at gaining an insight into adults' attitudes to cyberbullying and its prevention using technology, and this was achieved. Armed with this information, the researcher was

able to embark on the next phase of the research study, which was to understand young people's perspective of the same topic.

4.3 Understanding Young People's Attitudes to Cyberbullying and its Prevention

As adolescents are the primary target audience for the proposed mobile app, their views are critical to the success of the app. After adequately considering the sensitive nature of the topic being studied and the age of potential participants, it was felt that interviews were the more appropriate vehicle to engage with the young stakeholders. In a one-on-one interview, participants can freely discuss their experience and views on cyberbullying without peer influence or judgement. It provides a safe and confidential environment for the young participants to have an honest discussion about cyberbullying.

This phase of the research study was initially devised as a schools' project, and a number of local secondary schools were contacted to explore the possibility of collaborating on the project. One of the schools (a co-educational comprehensive school) agreed to take part in the study. Ethics approval was subsequently granted by the University Ethics committee to conduct the school project and, specifically, for the researcher to conduct individual interviews (with the school's counsellor in attendance) with students from Year 7 upwards to gain an insight into their opinions on cyberbullying and how the proposed mobile app could assist in its prevention and mitigation. Unfortunately, despite several fruitful meetings with the school authorities and favourable engagements with students by the researcher in the form of giving talks at school assemblies and seminars about cyberbullying and the research study as part of a staged recruitment process, progress slowed and eventually halted due to ultimate non-engagement on the part of the school. In order to progress the research, and mindful of the delay this had caused to the research, a decision was then made to, instead, recruit first-year Aston University undergraduates as the study participants. As young adults, many of whom just recently graduated from secondary school, first-year undergraduates are the closest in age to secondary school students amongst the undergraduate student population. Their views on cyberbullying are therefore likely to be representative of those of the initially targeted secondary schools students. A new ethics application was submitted to conduct

individual interviews and participatory design sessions with first-year undergraduates which was granted by the University Ethics Committee (see Appendix B.8).

A total of 25 interviews were conducted in the period December 2017 - July 2018, the findings of which are discussed in the following sections. Participant recruitment was performed via research invitation posters (see Appendix B.9) placed in various locations on the University campus and also via invitation emails (see Appendix B.10) sent to first-year undergraduates. Due to the University's data protection policies, the students' email addresses were not shared with the researcher; instead, the invitation email was forwarded to the research liaison of the University's Schools. The invitation emails were then sent by each School to all of their first-year undergraduates. As a result of this arrangement, the exact number of students emailed is not known; data provided by the University's Planning and Student Management Information unit indicated that 6065 students were admitted into the first year of the University in 2017 and so it is anticipated, therefore, that approximately this many students received the research invitation. The invitation email included a link to an online pre-study questionnaire (see Appendix B.11) that gathered demographic and social media usage information from respondents and was used to determine their suitability for the research interviews.

4.3.1 Pre-Study Questionnaire

Seventy-two respondents completed the online pre-study questionnaire, of which 51 identified as female (70.8%), 20 as male (27.8%), and one respondent preferred not to say. Forty-three respondents (59.7%) were aged 17 – 21 years, 26 (36.1%) were aged 22 – 24 years, and the rest were at least 25 years old. As illustrated in FIGURE 4.1, WhatsApp was the most popular social media/messaging platform used amongst respondents, followed by Instagram and Facebook. The popularity of Instagram amongst young people has also been highlighted by a study conducted by the UK's Office of Communication (Ofcom Research, 2019) where it was found that, while Facebook remains the social media platform with the highest number of registered users, this is mainly due to a high number of legacy user accounts and that adolescents and young adults in the UK are more likely to use Instagram, WhatsApp and Snapchat.

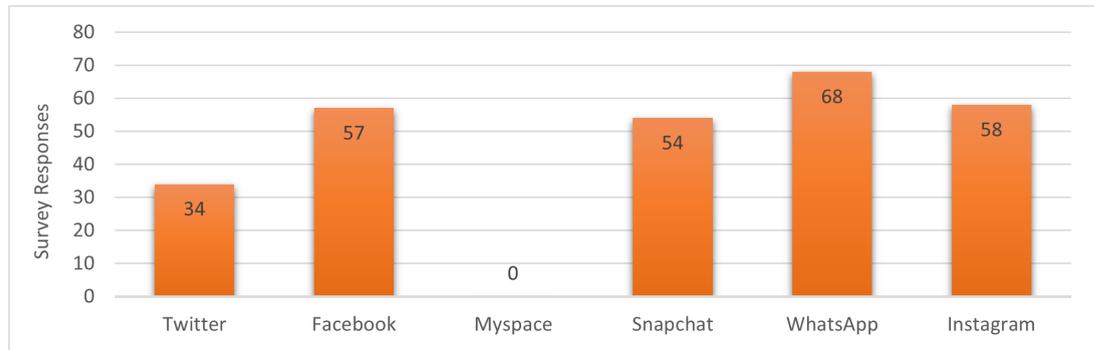


FIGURE 4.1: Social media platforms used by respondents.

Many of the adults in the focus group study shared the view that young people spend a significant amount of time on social media, and this opinion was confirmed given that three-quarters of the students surveyed indicated that they check their social media accounts several times a day (see FIGURE 4.2). Mobile phones' dominance as a means of accessing social media platforms and the Internet at large was also affirmed, with nearly all respondents selecting mobile phones as a means of accessing social media platforms (see FIGURE 4.3). This is in concert with Ofcom's findings (Ofcom Research, 2019) that 75% of the total time spent online in the UK is via mobile devices.

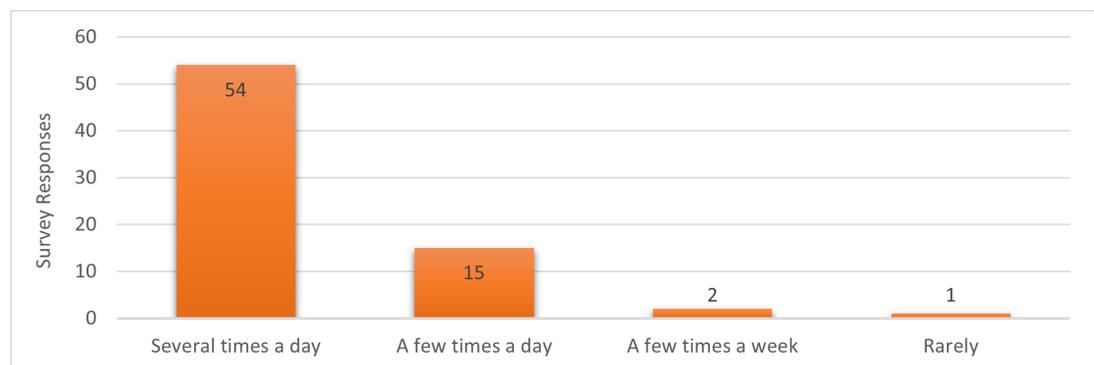


FIGURE 4.2: Frequency of social media interaction.

As discussed in Chapter 2, variations observed in reported cyberbullying victimisation and offending rates can sometimes be attributed to the timescale across which young people are asked to reflect in terms of their cyberbullying experience; usually, the longer the timeline, the higher the cyberbullying prevalence. Of the 41 respondents that admitted being previously bullied online or were unsure (see FIGURE 4.4), 31 of them said that the abuse occurred over a year ago, and the rest experienced online bullying within the last 3 - 12 months as shown in FIGURE 4.5.

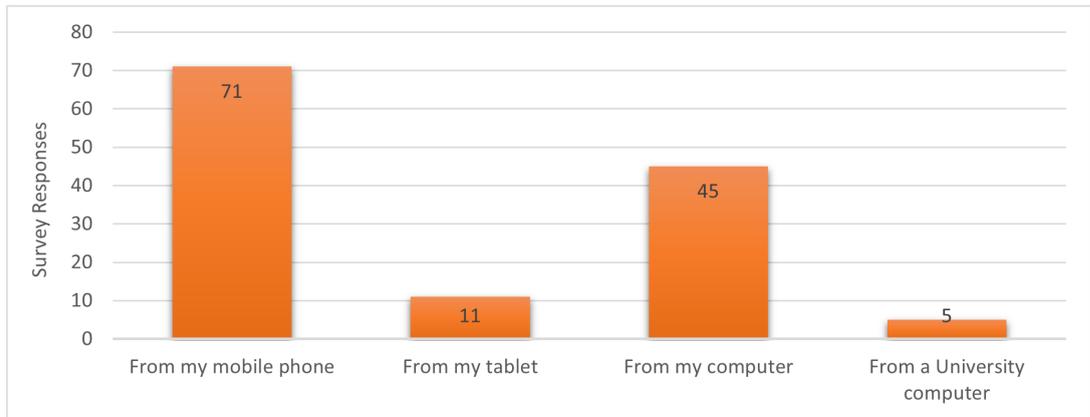


FIGURE 4.3: Means of accessing social media.

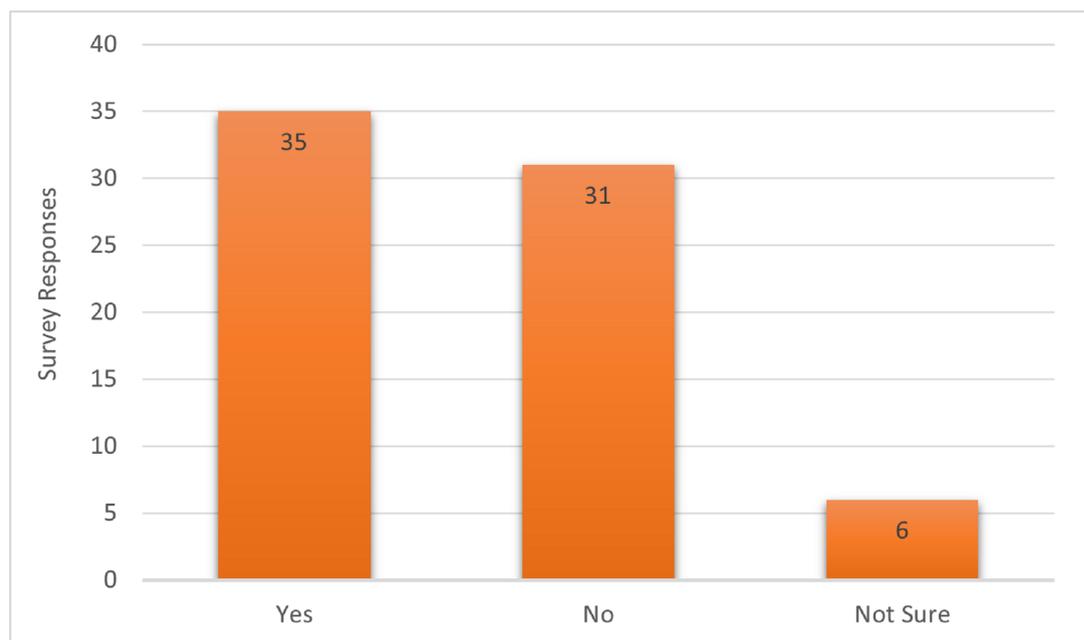


FIGURE 4.4: Number of participants that have been cyberbullied.

When queried on the specifics of how they had been bullied, the majority had had offensive comments posted about them online; this was followed by being abused via offensive text (SMS) messages and the sharing of inappropriate pictures online (see FIGURE 4.6).

Ignoring cyberbullies was the typical response of many victims (51.4%), but many respondents also said they did not know what to do (43.2%) while some were too scared or upset to do anything (21.6%). None of the cyberbullying victims in the survey reported the abuse to the school authorities, and only a handful confided to their parents or guardians. Similar to the findings of Bauman (2010) which noted that students rarely report cyberbullying incidents to teachers, the survey discovered that the majority of cyberbullying victims were passive in their response to being bullied and would rather

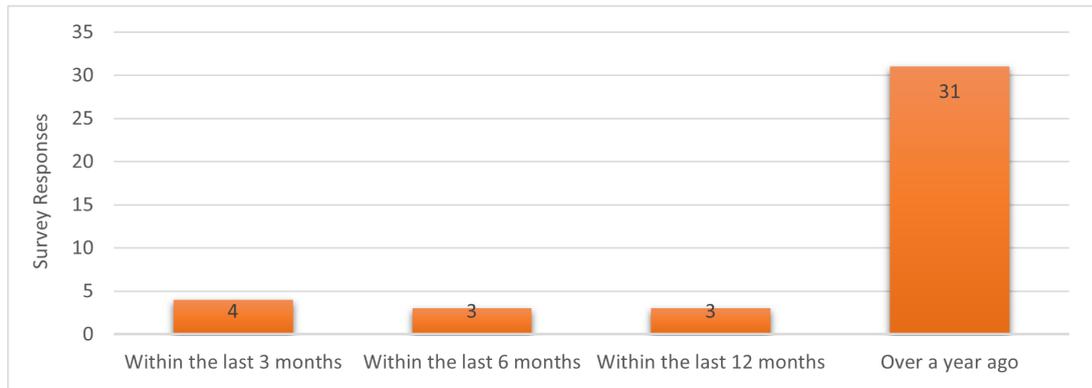


FIGURE 4.5: Period when the cyberbullying incident occurred.

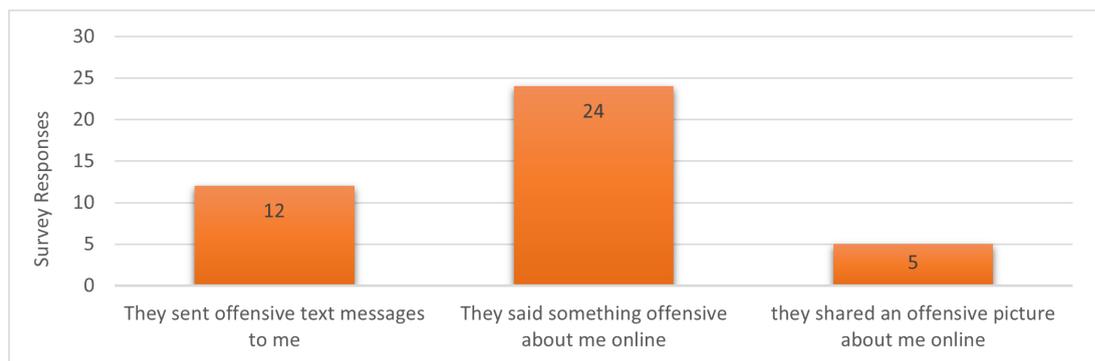


FIGURE 4.6: Type of online abuse experienced.

endure the abuse than report it to teachers or parents, as illustrated in see FIGURE 4.7. This has significant implications for cyberbullying prevention; rather than hoping that young people will confide in them and then being reactive, adults should actively engage with young people and create an environment where young people feel safe enough to open up about online abuse. Not all cyberbullying victims reacted submissively: about a fifth of them confronted the cyberbullies and told them to stop, while a handful retaliated by sending offensive messages back to the bullies.

Confiding in a friend also appeared to be a common reaction to being abused online. When considered alongside the fact that 13 out of the 40 respondents that had witnessed a friend being cyberbullied (see FIGURE 4.8) reported it to a teacher (see FIGURE 4.9), it can be seen that the support of friends is a critical aspect in combating cyberbullying. Encouraging strong positive relationships with peers should, therefore, be included as a crucial component of cyberbullying mitigation. To emphasise this further, the survey showed that more participants reported the abuse of a friend to a teacher than to another friend.

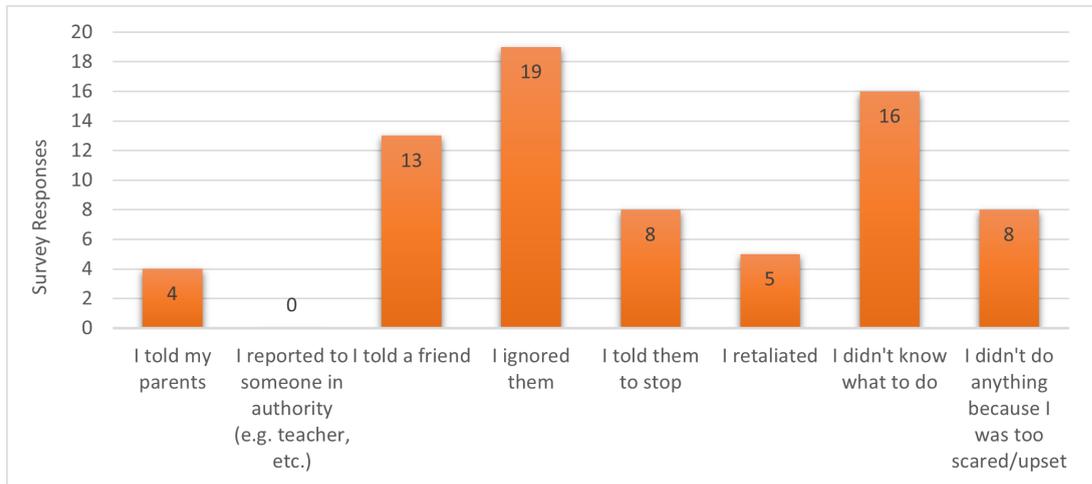


FIGURE 4.7: Respondents' response to cyberbullying when they are the victim.

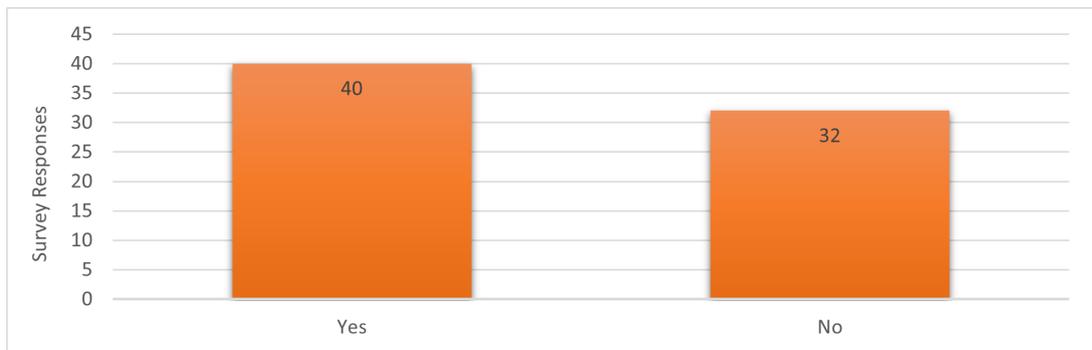


FIGURE 4.8: Number of respondents that have witnessed a friend being cyberbullied.

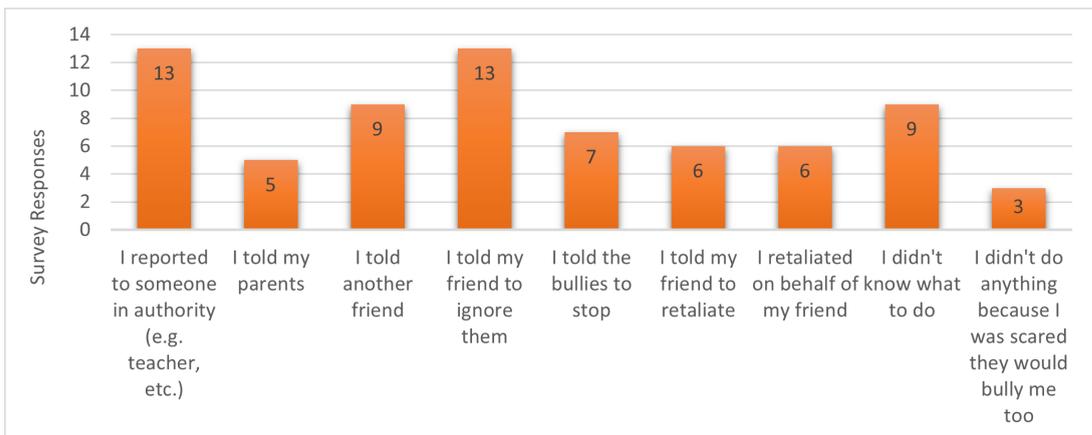


FIGURE 4.9: Respondents' response to cyberbullying when a friend is the victim.

Again, ignoring cyberbullies was a common tactic as respondents were as likely to advise their friends to ignore the bullies as they were to report the bullying to a teacher. Retaliating or advising the victim to retaliate were found to be the fourth most common tactics, preferred ahead of reporting to parents or doing nothing out of fear of being targeted as well. As discussed in Chapter 2, the low cyberbullying offending rate

recorded by many studies can be attributed to participants' desire to provide socially desirable answers (Hinduja and Patchin, 2013). This could be the case for this survey as well, as only three students admitted bullying someone online, and they did this by sending offensive SMS messages or posting something distasteful about the victim online. Finally, with regards the likelihood of using a cyberbullying prevention app like the one proposed, 69.6% of the students affirmed that they would use such an app.

4.3.2 Interviewees

Sixty-nine survey respondents indicated an interest in being contacted by the researcher to engage in follow-up individual interviews. The students were contacted via email and provided additional information about the research interviews in the form of a dedicated Participant Information Sheet (see Appendix B.12). Of these, forty-one students agreed to be interviewed, and interview sessions were arranged with all students at a mutually convenient time. Sixteen students, however, did not attend the original interviews nor re-scheduled sessions. In all, 25 individual interviews were conducted. The interviews took place on the University campus in small meeting rooms and study rooms to ensure privacy and limit the possibility of disturbance. Drinks and snacks were provided by the researcher. The interviews were structured as open discussions so that interviewees were at ease and the first 5-10 minutes were spent discussing common interests to build rapport between the researcher and the interviewee. A sample of the interview guide is presented in Appendix B.13. There were thirteen female, ten male, one trans female and one non-binary interviewee from the ethnic backgrounds summarised in TABLE 4.3. Twenty-two participants were aged 17 – 21 years, two were in the 22 – 24 years age range, and one student was 25 years or older. All participants consented to the interviews via the provided consent forms (see Appendix B.14 for a sample) and received a £10 amazon voucher for participation.

Ethnic Group	Number of Interviewees
Asian or Asian British	7
Black, African, Black British or Caribbean	5
Mixed or multiple ethnic groups	1
White	12

TABLE 4.3: Summary of interviewees' ethnic groups.

4.3.3 Emergent Themes

The interviews were audio-recorded, transcribed verbatim and stored in Nvivo. Similar to the analysis of the focus group data, an inductive analysis process was performed using the procedure established in Section 4.2.2. The eight key themes uncovered by the analysis are discussed in the following subsections, and the coding tables are presented in Appendix B.15.

4.3.3.1 Cyberbullying Occurrence Intensifies in Early Teenhood and Extends into the Late Teens

Williams and Guerra (2007) and Tokunaga (2010) suggested a curvilinear relationship exists between victims' age and victimisation rate in young people. They theorised that cyberbullying typically starts amongst children aged 10 – 11 years, peaking in ages 13 and 14 years, and declines rapidly at 14 years and above. The cause of this curvilinear relationship was, however, not expatiated on by the aforementioned studies. While the study findings here do support the possibility of a curvilinear relationship, the accounts of interviewees indicate that the nature of the decline in older teens is less steep than suggested as inferred from the comments below:

"The mean, malicious stuff didn't continue past year 12".

"There were some [incidents] in sixth form. Not as much as before but still quite a few. Some girls posted some stuff online about another girl, and she and her friends beat them up. It was quite a big deal at the time I think".

"I think there was a big age thing about it. I think a lot of it was start of secondary schools, I'm thinking year seven, year eight or toward year 10, 11".

"We were in sixth form when they created those pages to rate girls on their 'hotness'".

"My first year of sixth form, I was harassed by an ex and her family on Facebook Messenger quite badly".

"Thankfully, it stopped when we got older [...] most of the people that I've heard about, they stopped taking it serious[ly] because then we were like around 16, 17. They never took it seriously, so I didn't either. Online stuff didn't affect us as much as the younger generation. I think it was bigger for year 9 or 10".

It appears that as young people enter a transitional period in the final years of secondary school, their focus shifts to other concerns outside of secondary school like higher education and employment. This theory is substantiated by one of the student's observation:

"To be honest, I think when we got to sixth form, people couldn't be bothered. We were more concerned about college and Uni".

Another noted that:

"As we got older and obviously became more aware of things like mental health. I think people just sort of tone things down a bit. There was less of the cyberbullying definitely".

This study, therefore, identifies a shift in priorities and the awareness of the negative consequences of cyberbullying that comes with maturity as the two leading causes for this reduction in cyberbullying activities as children grow older.

Contrary to studies such as those by Olweus (2012) and Olweus and Limber (2018), the young people interviewed considered the impact of cyberbullying worse than that of physical bullying. Many of the students interviewed were very vocal about their opinions regarding this as exemplified by the following quotes:

"The effect that one person's comment can have on you, it can be huge really. A lot of people could end up killing themselves regarding it".

"Cyberbullying is, if anything, I find it more of a serious problem [than physical bullying] because it's almost undetectable at the moment".

"Physical bullying and all that is bad, but there's a lot of people who underestimate the harm that cyberbullying can bring."

This higher impact of cyberbullying (compared to physical bullying) as perceived by the students is primarily linked to the pervasive nature of the Internet and its ability to reach a broader audience. One student noted that:

"The thing is that on the internet, everyone can see it. Normal bullying, there's probably about ten people all around you. You would see what's going on but, on the Internet, literally, the entire world can see".

This difference in the potential number of witnesses highlights a crucial distinction between physical and online bullying and one that is often ignored by studies such as that by Olweus and Limber (2018). The possibility of an online bullying act being witnessed by such a large number of people and also being electronically preserved indefinitely can increase its potency and, consequently, the impact on victims. Smith (2012) made similar conclusions in disputing Olweus (2012) views on the reduced importance of cyberbullying compared to physical bullying. Equally as discovered in the literature (Fahy *et al.*, 2016; Kim *et al.*, 2017; Khine *et al.*, 2020; Martínez-Monteagudo *et al.*, 2020), the study uncovered similar concerns about the mental health dangers of cyberbullying for young people as illustrated by the following statements:

“My cyberbullying didn’t start until I was about 11. It was when I first got a phone, and at the time, in year six, everyone seemed to have Facebook [...] for me, my mom always took me out of school when I was 11 because of the constant abuse. It was pretty bad [...] that was when my mental health issue started because that gave me major anxiety. I didn’t want to go to school. My mom had to take me into school and pick me up from the reception. I was only 11. That’s pretty bad to be too scared to go to school”.

“I’ve had different friends saying they have been bullied online to an extent where it’s kind of overtaken their lives”.

Additionally, the study found these concerns to be well-justified as observed in the tragic case narrated by one of the participants:

“[...] he’s not really a close friend, sort of someone I knew. He was in year 11. On Facebook, there were allegations going on about him. That he was gay and that he had sexually assaulted someone. It sorts of spiralled out of control; there was a lot of aggression directed towards him, particularly on Facebook. He ended up killing himself. He hung himself in a tree in the local park”.

While suicide ideation as a consequence of cyberbullying has been identified in the literature (Iranzo *et al.*, 2019; Brailovskaia *et al.*, 2018), this study, unfortunately, substantiated this claim highlighting that the disastrous conclusion of such a horrific association is, unfortunately, a possibility. Out of the twenty-five students interviewed, twenty-four (96%) had experienced cyberbullying as victims, bystanders or perpetrators. This staggering statistic implies a prevalence rate much higher than suggested in the literature. As one student put it:

“I think everyone knows someone probably, who has been cyberbullied or they may have been bullied themselves”.

4.3.3.2 Appearance and Identity are Common Bullying Themes

Physical appearance, sexuality, race, culture and ethnicity were identified as common cyberbullying themes in the studies of Dinakar *et al.* (2011) and Dadvar *et al.* (2012a), and this became evident in the students' responses as illustrated by the following quotes:

"We were just debating, and I let it slip out that, 'Yes, I'm a mixed-race person.' So after I said I was a mixed-race person that's when the abuse came in [...] It just seemed strange because suddenly, okay, we're debating this thing, but then suddenly you find out that I am mixed-race and suddenly the debate is over. This is when you start hurling the abuse".

"[...] it was always very subtle, but you could tell that she was being unkind to maliciously hurt my friend. It was really sad to see. It was more picking on and making comments about her weight or bad hair or things like that. It was more subtle. I think that one is more dangerous because it's subtle."

"It was Facebook. They were just calling her names, and there was a picture of her. She doesn't look really great in it. People thought she looked ugly basically and so they made fun of it".

"They said he was gay and that he had sexually assaulted someone".

"[...] repeated comments about things like my weight, my accent, my appearance".

It would appear that anything that makes a young person different in any way could be potentially used by bullies to offend. This could be physical features, accent, race, culture, identity and medical conditions. The typical student profiles associated with traditional bullying are also quite fluid in cyberbullying; for example, popular students who are generally unlikely to be bullying victims have been subjected to online abuse. A participant recounted a situation involving a popular student in her former school:

"I knew this girl, she quit the school, that was probably because of all the cyberbullying she got, I don't even get why they were picking on her, she's quite pretty I think that's why, weird right".

Another student with cerebral palsy narrated his experience being bullied about his condition:

“You know Breaking Bad [a TV series]. Walter White’s son, Junior has cerebral palsy in the series and I have that, so they used to call me Junior”.

He went on to say that his abusers later apologised to him in the following year and became his friends. This further substantiates the study’s finding that as children grow older and become more aware of the impact of their actions, their proclivity for cyberbullying offending reduces. The implication, therefore, is that reflective tools and activities that encourage empathy amongst young people (as suggested by the adult stakeholders) are key strategies for preventing cyberbullying.

4.3.3.3 Cyberbullying on Facebook and Twitter is More Public Compared to Snapchat and Instagram, Where it is More Personal and Targeted

As revealed by the pre-study questionnaire, WhatsApp, Facebook, Twitter, Instagram and Snapchat were the most frequently used social media and messaging platforms by young people. Although WhatsApp was the most popular messaging application used, none of the participants experienced or witnessed any form of online abuse on the platform. This is likely because, as reported by the students, WhatsApp is mainly used to communicate with close friends and family. Facebook, however, was frequently cited as the platform on which many of the cyberbullying incidents occurred. With Facebook accounting for 58% of the monthly Internet traffic for social networking (Statista, 2020b) and as many as 80% of the pre-study survey respondents reporting that they use Facebook, it is not surprising that a substantial proportion of the online abuse experienced by interviewees happened on Facebook, as typified by statements such as:

“There was a Facebook page created by some kids in school [. . .] I found my name written multiple times, but mostly it was nice things. Then there was that one time when someone didn’t say something quite as nice”.

“A friend of mine had a Facebook account made for him; someone else made it under his name they were pretending to be him saying things like ‘Oh, I’m gay and such an idiot.”

In addition to Facebook, Twitter, Instagram, and Snapchat are some of the other social media platforms where interviewees had experienced or witnessed cyberbullying. The manner of the abuse on these networks varied and included hacking:

“When I was in sixth form last year. It was on Instagram and Snapchat. What happened was I think, so a person logged into my friend’s Instagram account and saw our messages and then they obviously used that as information against me. They moved that to Snapchat so everyone could see it, within a few minutes a lot of people could see the Instagram private messaging”.

stalking:

“She [the tormentor] would also try and get to her through other people. I don’t think she was trying to be nasty. She was just obsessed with her like some people are. She was basically stalking her on Instagram and Twitter but she kind of forced her to close down a lot of her online life as a result”.

and sending offensive messages:

“Snapchat is a bad one I think for cyberbullying because it’s very personal and once that snap’s gone, it’s gone”.

“Twitter is quite vile. It is not a nice site at all. Twitter people are just not nice I have just like posted inspirational stuff, and random people will start sending you abuse for no reason”.

“They were sending me DMs [Direct Messages] saying that they were going to post all sorts of horrific untrue accusations about me on social media and things like that in order to discredit me”.

The different types of cyberbullying experienced or witnessed by participants coincide with the categorisation proposed by Bauman (2015), highlighting the various manners in which cyberbullying can be manifested. A crucial distinction was, however, observed by the study in the way cyberbullying is perpetrated on Facebook and Twitter compared to Instagram and Snapchat. The types of online abuse experienced and witnessed by participants on Facebook and Twitter can be likened to publicity stunts designed to offend the victim in a very public manner. These types of abuse are often perpetrated by creating a page or a fake account as a means of ridiculing and causing distress for the victim. The comments from participants indicate that the more people who are aware of a cyberbullying incident, the higher the distress caused by the incident, thus attributing public forms of abuse as the more severe form of cyberbullying. This is in concert with the findings of other studies like those of Nocentini *et al.* (2010), Slonje and Smith (2008) and Sticca and Perren (2013) which similarly found public forms of cyberbullying the most distressing for victims. The dangers of this type of bullying have been highlighted in the media (Davis *et al.*, 2012;

Evans, 2018) thankfully these forms of online abuse are now quickly taken down by social media platforms once discovered (Cheng *et al.*, 2019).

The other type of cyberbullying, which is more frequent on instant messaging-oriented online social networks like Snapchat and Instagram, is targeted messages sent directly to the victim typically from anonymous accounts. While not publicly visible and thus only seen by the victims, they represent an intimate form of abuse and can contribute to a feeling of isolation for the victim. A student that admitted bullying an ex-partner online via Instagram said this:

“She was utterly bombarded by me like five or six times a day, all these things because I used the Internet effectively as my tool to have that effect on her”.

By understanding how cyberbullying is perpetrated on different social media platforms, cyberbullying prevention tools such as the proposed mobile app can be specifically designed to target these forms of cyberbullying on the relevant platform, thereby improving their effectiveness. As discussed in Chapter 2, this research favours a redefinition of the repeatability criteria often used to qualify cyberbullying to include single incidents where offensive material is publicly shared and made available for repeated viewing. The data from this study certainly supports this. Such is the speed of propagation of social media that content available for mere seconds can be instantly captured and shared to a multitude of people. As noted by one participant:

“I went to an all-girls school; it was very frequent. In my school, when there’s an argument, people start exposing people and different things on social media. They posted indecent pictures of a girl [...] Everyone saw it. She just said that it wasn’t her, but it was. She didn’t do anything other than that. It was just the one time. She took it down. Everyone already saw it; everyone had screenshotted it and all that. It was too late; the damage had already been done. They will put it as their display picture so all their contacts could see it”.

4.3.3.4 Cyberbullies are Often Known to Their Victims

Only two participants admitted to bullying other people online. In both instances, the bullies were male, and the victims were people known to them. One of them electronically harassed his ex-girlfriend for several months: he said at the time it never occurred to him that what he was doing amounted to abuse. In his words:

*"I was probably a really nasty person to, I guess, one of my ex-girlfriends. I kind of did that without realising it. I was discovering how useful the Internet was and became enamoured with the girl and effectively tracked down everything she had online and tried to get in touch with her by all those means. She asked me to leave her alone, but I didn't. I wasn't trying necessarily to do anything like bullying, [...] it really wasn't until she screamed at me to leave her alone that I realised what an a*s I have been."*

As stated by Smith *et al.* (2008), the absence of immediate visual or aural feedback from the victim can often hide from an offender the impact of their actions, and this certainly appears to be the case with the two interviewees in this study. In narrating his own experience, the other student stated:

"What did I do? I was probably 14-ish. They were my friend weirdly enough. We had split apart a little bit. I found his email address. I thought it'd be really funny to, this is terrible by the way, especially in today's means, but I sent loads of emails. God. Like, say I was a random person. That I was going to go to his house. I was going to blow him up. Really bad. Awful, awful, awful stuff. I don't know why I did it. I did it in order to rile him a little bit because we'd fallen out. I thought it was funny. The intent was to upset him but with the effect that we would laugh about it"

"I don't know what the situation was in my head, I don't know, but along the lines of make a joke about it and then they would stop, and we would be friends again"

In proposing their "revenge of the nerds" hypothesis, Ybarra and Mitchell (2004) suggested that online bullies may have previously been victims of physical bullying who then use the electronic medium to retaliate against their tormentors; this was not the case for the two cyberbullies in this study, albeit it is recognised that perpetrators only accounted for two interviewees. For the cyberbullying victims, many reported that they knew their tormentors who were former boyfriends and girlfriends, classmates and "people from school":

"I had a MySpace account back when it was cool many years ago, and I moved to Facebook at some point when Facebook came in. When I left MySpace, that snapshot of my life, I had certain things on my mind. I said some stuff on there, I forgot about the page and never went back to it. Then, about a couple of years later, a friend of mine had obviously gone and found that and used it against me [...] he threatened to copy it to Facebook as a screenshot"

"I've been slightly cyberbullied, but it was by an ex-boyfriend. It was on a couple of different platforms. It was Facebook Messenger and Tumblr. It wasn't exactly directed at me, just on posts which I could see. Being unkind about me for

everyone to see. They didn't tag me, but there were a couple of times when he said my name. I knew it was about me because it was specific to things that happened".

"My first year of sixth form, I was harassed by an ex and her family on Facebook Messenger quite badly".

In some cases, the abuse was extended to other members of the victim's family:

"My ex and some of her friends. It was just repeated comments about things like my weight, my accent, my appearance. If I blocked them, then they'd create new accounts. It wasn't just on social media; it's also over phone calls and text messages and things like that. They were sending messages to my parents. My little brother".

The quotes below from two interviewees who were severely abused online represents a rather succinct summation of the most likely culprits of cyberbullying:

"I didn't experience cyberbullying from people I don't actually know. They were always people I know. Always".

"[. . .] they were in my class. it was awful, what they did say. I did know it was them, there wasn't anonymity involved".

While the perceived severity of anonymous and non-anonymous cyberbullying have been studied and anonymous cyberbullying identified as the more severe of the two (Vandebosch and Cleemput, 2008; Smith *et al.*, 2008; Dooley *et al.*, 2009; Barlett, 2015), little is known about the prevalence of both forms of cyberbullying. Hence, this study's finding that the majority of the victims knew their perpetrators contributes new knowledge to the literature and exposes an area of cyberbullying research that has not been well explored.

4.3.3.5 Fear of Reprisals and Inadequate Responses Discourages Cyberbullying Reporting

"If you tell your parents, they tell the teachers. Then obviously the teachers tell the bullies, and it creates this sort of situation. I know if you look at it on paper it makes sense because when you're telling a person in authority about it, their idea is although they will be able to help you, the fact that you don't live in the kind of environment where the authority is always there to help means that in the times when they are not there, then the retribution for what you did in the first place makes it worse, makes it not really worth telling at all in the first place".

The above quote summarises young people's attitude to reporting cyberbullying to an authoritative adult. The data from the interviews support similar findings from studies such as those by Gao *et al.* (2016) and Sticca and Perren (2013) and indicates that young people do not report cyberbullying to adults out of fear that adults may overreact or that adults lack the relevant knowledge about social media and are thus ill-equipped to deal with cyberbullying. In discussing their reluctance to involve school authorities, some of the interviewees said:

"People who haven't witnessed it or haven't experienced it don't really know much about it. Especially parents who haven't been through when they were younger, they are just like, 'It's normal, just ignore it'"

"We never did really talked to the teachers. You go to the teachers, what are they going to do? As soon as they tell these people off, it's just going to start right back up again and maybe worse because now they actively hate you at the same time"

For the majority, however, the reluctance to report cyberbullying to teachers is out of fear of reprisals from the bullies if discovered, as well as the general stigma associated with reporting and being labelled a "grass" or "snitch" by other students. Some of the students recounted:

"I didn't want to go to school because I know there'd be a chance they'd tell my parents, and also there's definitely an element of not wanting to be a grass. I typically tell my friends. Not that they could do much about it because they were also typically being cyberbullied at the same time"

"I typically do report things that I think have been particularly harmful. If it looks like a lot of stuff could be focused at me if I intervened then I will report anonymously. I just don't want to have that sort of stuff directed towards me"

Not all the students were, however, averse to reporting cyberbullying. The results for those that did were, however, mixed. While some achieved positive outcomes:

"My parents got involved, they essentially threaten legal action against them if they don't stop this [. . .] it was particularly bad, but it did get sorted"

"I've had friends who've been affected badly because they've received messages like harassment. The harassment was in a sexual nature, so we got the evidence, and we went to the school head to report, and the guy got kicked out"

“I reported them, and then I blocked them because I didn’t want to see anything else, then I got my friends to also report it, but it got taken down in like hours, so that was good, but I think we were lucky because some other times, Facebook didn’t respond at all.”

For others, however, the experience was not as positive. One participant said:

“I told the assistant principal of my sixth form, and then she said that hacking is impossible of an Instagram account [...] and she was like, It’s not hacking, it’s a fight. Get over it’. I find it quite hilarious, but I was done with that school within a month, this happened last May, and I was finishing June so, I knew I was out of there. This is the assistant principal, I went to her. I was crying, she shouted, “Get over it. It’s normal”.

There is undoubtedly a sense of helplessness on the part of the victims. They feel that since the abuse is not physical, there is very little the school can do:

“When it’s at home with your phone your school can’t do anything about it so why bother to tell teachers”.

“Teachers don’t know how to deal with it. The teachers aren’t educated properly about what it’s like to grow up in the time like us. Some of them want to help; some of them don’t care. The teachers that really do care about pupils really do stand out, but most of them just don’t get that. They don’t think that they can help. They see pupils getting bullied or other stuff, and there’s stuff happening, and they think, there’s nothing I can do about it, I’m just a teacher”.

“We don’t go to our teachers asking for support, so we get bullied more every day, and this keeps going on, there is a certain mental pressure on you. For example, I felt sick because I was very scared, it reaches a point when you can’t take it anymore, you have to share it with someone. If you’re not being able to do that, it just makes you devastated, frustrated”.

4.3.3.6 School Should Intensify Cyberbullying Prevention Efforts

When questioned about their schools’ strategies for mitigating and preventing cyberbullying, the general opinion of the students was that schools need to do more than *“a few assemblies”* and *“the odd poster”*. They said:

“We probably had a few talks about depression, speaking out, but I’m not so sure about cyberbullying specifically”.

“We did have a few assemblies about cyberbullying in our citizenship lessons; we had some documentary, kind of an informative film about cyberbullying”.

"I never knew the school to ever intervene with anything, cyberbullying or otherwise. Not at all. There were no posters or anything. Nothing. Not that I remember. If there was, it didn't stand out enough for me to see it".

"Sometimes I feel like schools are afraid to deal with the bad children because the parents are just as bad".

"In year seven, we were told that, if they wanted to, they could look at our social media records with the police. I don't think that deterred anyone".

With the students having similar opinions to the focus group adults about schools' efforts to prevent cyberbullying, it is clear that more needs to be done by schools to support both students and parents in terms of cyberbullying via education and raising awareness about cyberbullying. It is therefore critical that schools adopt a zero-tolerance attitude to cyberbullying, enabling an environment where students feel safe enough to report cyberbullying without fear of reprisals from the bullies. It is also vital that students are adequately reassured that their complaints will be fairly heard and treated seriously – i.e., that because the abuse occurred online does not mean it will be minimalised and that they cannot rely on the school for support.

4.3.3.7 Relevant Advice and Punitive Actions are the Critical Features for the Proposed App

In discussing the proposed app, many commonalities were observed in the students' suggestions on how the app can assist cyberbullying victims. Many felt that it was important for the app to support cyberbullying victims with relevant advice on how to tackle the situation. They were, however, very critical of existing advice on dealing with cyberbullying. One student noted:

"I think if advice were to be given and used on the app, it would be important whoever is writing the advice. Wherever it's coming from, they understand not just the victim's mentality but the bully's mentality as well because they are finding that the people doing the bullying are often raised in places where they have had violence and aggression towards them, they pass it on. I've seen advice, and they said things like - Just say no. Firmly say to your bully 'Please stop'. That's just not going to work. That's written by somebody who has no idea. Has no idea why it's happening, how it's happening and how to stop it".

The quality of the advice and the manner in which it is provided was therefore identified as a key aspect of delivering this feature. The students favoured advice that was relevant

and tailored to specific situations. Interestingly, while the students were quick to propose different types of sanctions that can be applied against online bullies by the app, unlike the adult stakeholders, they did not suggest features to encourage positive attitudes and empathy amongst potential users of the app. When the use of reflective interfaces was suggested by the researcher, only a small number (four) were enthusiastic about this. Of these four users, three have witnessed close friends being bullied, and one has been a victim of cyberbullying. The remaining students were of the opinion that reflective interfaces would not deter a bully from offending. One said:

"I don't think there's any way of preventing cyberbullying. At the end of the day, no matter what you're saying to people, teach people, they're always going to go do their own thing".

The two students that have bullied others online before were of similar opinions. They said it is doubtful that a reflective interface would have dissuaded them from performing the actions as they did not feel they were doing anything wrong at the time. After the researcher clarified that the typical use for such reflective interfaces would be to reduce instances of unintentional abuse or the use of inappropriate language during an emotional outburst or in the heat of the moment, the students were more appreciative of the feature's potential. As the interviews progressed and explored specific ways that the app can be used to reduce online abuse, the students were asked the question: how do you defend yourself against a cyberbully? The participants' key recommendation can be summarised as to "block and ignore" online bullies. In narrating their experience dealing with online abuse, some of the students said:

"I just started ignoring it and stopped responding. I stopped letting it affect me".

"What she did, she just literally left the account and made a new one. Made one that only her friends knew her from. That went quite well".

"I think she blocked her in the end".

"I think my mom actually found out. She saw all insults on the messages, and she got quite upset about it. She told me to block him, and so I did".

They followed these up by suggesting that the app should implement features that automatically sanction abusive users:

"If it can delete and block stuff, then that's good if you are being bullied [...], but if you are a bully I don't think an app will make you think twice".

"I think developing an app that could protect them or maybe just help them sort things out, I think would be a noble cause. Blocking the bullies is very good".

"Getting them [bullies] banned will be good, I think. Because they are always creating new accounts so if there's a way to just keep banning the new accounts till they get the message".

4.3.3.8 Young People Would Rather Report Cyberbullying Anonymously than get Directly Involved

The influence of defenders (individuals that stand up to bullies in support of victims) on the possible outcomes of a cyberbullying incident has received some attention in the literature and empathy has been positively correlated to the likelihood of a witness defending the victim (Machackova and Pfetsch, 2016; DeSmet *et al.*, 2016). When questioned on what their initial reactions would be if they witnessed cyberbullying, the majority of participants said they would avoid a confrontation with the bullies but would consider reporting the incident if there were ways to do so anonymously:

"My instinct is to say yes because you would want to be the kind of person who would say yes. Without knowing the situation, it would be difficult I would think. Because, you know, what would you say? It might even spark it into being more aggressive, in all honesty".

"Is this going to come back and hurt me in a certain way? If It would be more like yes. I wouldn't. I'm not an online superhero. That's not me".

"I'm not sure I would to be honest. I like to think I would but the bystander factor, meaning people generally just tend to ignore things that aren't happening to them or aren't important to them. I like to think I would, but I'm not sure".

"If there was a way to report it that was anonymous or wouldn't get me involved, I would probably report it".

"I would definitely complain and go to some authority and inform them about the cyberbullying, but I won't engage them [the bullies] directly".

This implies that, in addition to empathy, self-preservation is another factor influencing young people's decision to intervene when witnessing cyberbullying. Not all students share this opinion, however, as a few students said they would take a more active role and speak out in defence of the victims or contact them to offer support as seen in their comments below:

"I would tell them to stop Yes. Obviously, I would be, kind of, scared but I would think that's the right thing to do, certainly, yes. I would tell them to stop".

"I like to think, like, I'd definitely want to do something, and I definitely do want to".

"If I thought it was a bit serious, it was getting a bit too much. I would personally message the victim".

4.3.4 Discussion

It is clear from the literature, and the studies conducted, that cyberbullying is a complex issue and one that requires multiple strategies to manage different situations. The interviews provided an enriched insight into cyberbullying from the perspective of young people. Whilst many of the findings reinforce those from the literature, the study also uncovered some contradictions and discoveries.

The notion that cyberbullying prevalence amongst young people reduces significantly after ages 14 – 15 years was found not to be the case as many of the study interviewees reported experiencing or witnessing cyberbullying up until they graduated from secondary school. Furthermore, while the students believed that there was a reduction in the overall occurrence of cyberbullying incidents during the period they were in sixth form, their accounts suggest this was more likely a steady decline than the steep reduction suggested in Tokunaga (2010) and Williams and Guerra (2007). The reduced but sustained cyberbullying activities witnessed during this period can be attributed to two opposing influences. Firstly, as they grow older, and their maturity increases, children become more aware of the negative impacts of cyberbullying. Additionally, as they prepare for life outside secondary school, there is a shift in their focus as they look ahead towards higher education and employment. These combine to reduce their penchant for engaging in cyberbullying activities resulting in a subsequent reduction in cyberbullying prevalence at these ages. In contrast, the increased access to mobile devices as well as improved knowledge and sophistication in the use of the Internet, coupled with reduced supervision provide older teenagers with more opportunities to engage in abusive behaviour online. The resultant effect of these two factors is the continued cyberbullying occurrence in late teenhood as observed by the study.

While a lot has been said about the anonymous nature of the Internet and the ability of cyberbullies to use this to torment their victims (Smith *et al.*, 2008) thereby increasing the

perceived severity of bullying incidents (Dooley *et al.*, 2009; Barlett, 2015), the majority of the interviewees who had been bullied online revealed that their abusers made little attempt to hide their identities. This is an interesting discovery that suggests a parallel narrative to that of the anonymous abuser typical in the literature. As to why an abuser would knowingly choose to allow their identities to be discovered, this is likely related to the motivations for bullying. Of the two participants that admitted bullying others online, one made concerted efforts to hide his identity to increase the level of distress to the victim. In discussing his motivations, the student said:

"[. . .] like, say I was a random person. That I was going to go to his house. I was going to blow him up. Really bad. Awful, awful, awful stuff. I don't know why I did it. I did it in order to rile him a little bit because we'd fallen out".

Based on these comments, it can be seen that anonymity was an intrinsic part of the attack, which was designed to make the victim paranoid about the possibility of a sudden attack from an unknown assailant. Furthermore, this incident also represents a form of online bullying that is largely ignored in literature, which is the threat of violence thus validating the inclusion of a threat label in the taxonomy of the cyberbullying dataset created in the previous chapter.

In the case of the other student, the motivation for harassing an ex-partner was to rekindle a broken romantic relationship. The student had convinced himself of the legitimacy of his actions, and therefore it was important for the victim to view the harassment as a declaration of continued romantic interest, hence the decision to reveal his identity. It was also discovered from the students' accounts that for some forms of cyberbullying where the intention is to publicly humiliate the victim, for example creating "rating" pages to rate physical appearances or fake profiles to spread malicious content, it is common for abusers to not hide their identities to recruit collaborators and demonstrate their lack of fear of reprisals from the victims. The motivations for online abusers' to reveal or conceal their identities is, therefore, an area of cyberbullying research that should be explored more. Cyberbullying victims also tended not to report the abuse to parents or teachers out of fear of reprisal from the bullies or because they were unconvinced of adults' abilities to protect them. This, unfortunately, creates a feeling of helplessness amongst victims which can cause further damage to their mental health as observed in the following students' accounts:

"I told the assistant principal of my sixth form [..], and she was like, 'It's not hacking, it's a fight. Get over it'. [...] This is the assistant principal, I went to her. I was crying, she shouted, "Get over it. It's normal".

"We don't go to our teachers asking for support, so we get bullied more every day, and this keeps going on, there is a certain mental pressure on you. For example, I felt sick because I was very scared, it reaches a point when you can't take it anymore, you have to share it with someone. If you're not being able to do that, it just makes you devastated, frustrated".

The students were very critical of their schools' policies on cyberbullying, and believe that much improvement is required in schools' efforts in mitigating and preventing cyberbullying. Their views in this regard are similar to those of the adult stakeholders, which suggests current cyberbullying prevention efforts by schools are not seen as being effective.

While empathy has been positively correlated with attitudes to defending cyberbullying victims (Cleemput *et al.*, 2014; Machackova and Pfetsch, 2016; DeSmet *et al.*, 2016), this study identified self-preservation as another factor influencing young people's decisions to intervene in cyberbullying situations. As is to be expected, the need for self-preservation is strongly linked to a young person's propensity to defend a cyberbullying victim. The reluctance of a witness to assume the defender role does not necessarily indicate a low level of empathy (as is often suggested in the literature); rather, another possibility is that self-preservation desires are more potent than the feelings of compassion towards the victims. Furthermore, the self-confidence possessed by the witness plays a significant role in the decision to intervene. The three students in the study who said they would intercede on behalf of the victims appeared very self-assured. They were very confident in their opinions about cyberbullying and its prevention, and this would have contributed to their inclination to become defenders. This finding implies that providing an environment where students can safely report cyberbullying incidents without fears of reprisal should be adopted by educational institutions as a fundamental aspect of the overall cyberbullying prevention strategy.

In suggesting desired features for the proposed mobile app, the students' suggestions were primarily focussed on punitive actions such as automatically blocking and reporting abusive users and deleting offensive messages. As would be expected given they are more familiar with technology in comparison to the adult stakeholders, the students were more specific when discussing their desired features for the proposed app. When

questioned on the level of performance expected of the proposed app, the majority said they would expect the app to detect at least half of the abusive messages accurately with many saying they would stop using the app if the performance drops below this level. About a fifth of the students, however, said that they would continue using the app even if the performance falls below this level. Four of these five students were former cyberbullying victims, and their past experiences as victims may have played a role in their desire to continue using the app as long as it provides some form of protection.

While all of the interviewees confirmed that their secondary schools used content filters and web monitoring software on the schools' computers and wireless networks, only a handful (three students) had used similar software on their mobile devices. For these students, the software was installed by the parents; two of them managed to disable the monitoring software themselves, and the third protested and got her parents to remove the software. The students were aged 14 -15 years at the time. Unsurprisingly, the students were very vocal about their objection to the notion of the proposed app featuring parental monitoring elements. Apart from this contradiction, many of the young people's desired features (see TABLE 4.4) overlap with those suggested by the (adult) focus group participants implying that both stakeholder groups identify similar and complementary needs for the proposed app.

The ages of the students interviewed (17 – 25 years) are higher than those of the secondary school students (11 – 17 years) initially planned as the study participants; as such, the level of maturity possessed by the first-year undergraduates would be absent in the secondary school students. This maturity is evident in their views on cyberbullying and how the proposed app can tackle it. Furthermore, the students' perspective is wholly retrospective as they recall incidents from their past. Eighty-eight per cent of the participants are 17 – 21 years and would have been secondary schools students within the last three year; as such, their views are still very much relevant and representative of a large section of current secondary school students. The maturity of the students also provides them with an objectivity that would be missing in younger secondary school students, and each interviewee provided the study with the benefits of their accumulated experience across different stages of secondary school and how cyberbullying affected them through these different stages. Additionally, it is doubtful that younger secondary school students will reveal as much about their experience as online bullies to the researcher in the presence of a member of the school staff (as planned in the original

Reflective, educational and empowerment	Punitive	Others
<ul style="list-style-type: none"> • Display daily motivational quotes • Provide access to relevant advice and help for cyberbullying victims 	<ul style="list-style-type: none"> • Block offensive users • Report offensive users to the social network • Automatically delete offensive messages 	<ul style="list-style-type: none"> • Allow 'whitelisting' of contacts • The app's design should not be too childish • Keep copies of deleted messages • Users should be able to review deleted messages • Permanently remove deleted messages after a configured period

TABLE 4.4: Young People's desired features for the proposed app.

study protocols) as some of the participants have. Whilst the study participants are older than those of other studies on cyberbullying, a key distinction between this study and the literature is that the current study is not investigating present cyberbullying prevalence amongst participants, rather it is exploring participants' views of cyberbullying as influenced by their historical experience with the phenomenon.

The researcher, therefore, does not believe that the age of participants has adversely impacted the study or any of its conclusions, rather the research has benefited from the objectivity introduced by the participants' age and experience.

4.3.5 Study Limitations

As previously noted, the original intention for the study was to conduct interviews with students of a local secondary school, but this strategy had to be revised due to circumstances outside the control of the researcher; given that recruiting one school to the study had proven difficult, the decision was made in the interests of progressing the research to refocus attention on Aston University's first-year undergraduate student cohort. The consequence of this is that, rather than interviewing students currently in

secondary school, the interviews were conducted with students that recently left secondary school. Given most of these undergraduates recently progressed from secondary school, it is felt that their experience of cyberbullying remains relevant and representative of those of secondary school students. Their current views on the subject may, however not be representative of secondary school students. Despite this, the insight gained from the interviews is critical to understanding how young people feel about cyberbullying, its impact on their lives and their opinions on how it can be prevented using technology. Furthermore, the sample contained young people who have experienced cyberbullying as victims, perpetrators and bystanders, providing a range of perspectives that is only possible via experience.

4.4 Summary

This chapter presents the findings from studies conducted to understand key stakeholders' views on cyberbullying and the ways via which the proposed mobile app can assist in its prevention.

While the profiles of stakeholders involved across the studies are vastly different, the studies uncovered substantial overlap in stakeholder opinions and desires, particularly in terms of the functionalities desired in the proposed mobile app. Though adults believed that there is adequate information available online about cyberbullying, they are of the opinion that the information can be improved by making it available in easily digestible formats. It was also suggested that schools should be more responsive in their approach to tackling reported incidents of cyberbullying and online abuse. To facilitate this, teachers would require access to specialised training and resources on managing cyberbullying amongst students. Social media companies were also criticised for their inability to adequately address the abuse that occurs on their various platforms.

Young people shared similar sentiments about social media companies and their approach to confronting online abuse. Their reluctance to report online abuse to parents and teachers could potentially be obscuring how pervasive cyberbullying truly is amongst young people. Encouraging cyberbullying victims to be more forthright in reporting the abuse may be challenging for schools, but, as discovered in this chapter, young people are more likely to report online abuse suffered by their friends; this suggests that this is

an area where schools have the potential to be more successful by enabling an environment where students can safely report abuse. It is also crucial that when online abuse is reported, it is treated seriously and given adequate attention by the school authorities. A situation like the one experienced by an interviewee whose complaints were ignored by the assistant principal is not acceptable.

The functionalities desired within the app by all stakeholders are very similar. Interestingly, features that facilitate restrictive supervision were not popular, even amongst adult participants. This suggests that adults are not in favour of app features that may reduce the app's acceptance amongst young people; instead, they prioritised reflective and protective features that encourage positive attitudes while keeping young people safe on social media at the same time. Raising awareness about cyberbullying and providing access to specialist support for cyberbullying victims were also identified as essential features for the cyberbullying prevention app.

In the next chapter, these findings are incorporated into the design process as part of the participatory design approach adopted for the design of the proposed mobile application.

Chapter 5: Participatory Design of the BullStop Mobile Application

5.1 Introduction

The previous chapter outlined the investigative activities undertaken to gain insight into stakeholders' perspectives on cyberbullying prevention, the outcome of which was an initial list of desired functionalities for the proposed mobile app. This chapter reports on the use of Participatory Design (PD) methods in the development of BullStop. The use of PD enhanced the development process in two key areas. The first is that by collaboratively creating design prototypes with a cohort of Aston University first-year undergraduate students as co-designers; the developed application is a manifestation of the features and functionalities desired by the target audience resulting in a final product that is acceptable to the target audience.

Secondly, the engagement of both adult and young stakeholders in defining the behaviour of the application is injecting human sensibilities into the design of an Artificial Intelligence (AI)-based system. Often, when AI-based systems are designed, the primary focus of the developers are the performance gains introduced by the use of AI and the impact of such a system on the human users are rarely given adequate consideration. The approach adopted by this research ensured that the impact of the decisions made by the system on humans are adequately considered and used as the primary driver for the design of the system.

Section 5.2 provides an overview of Participatory Design and how it facilitated creating an acceptable and impactful tool to assist cyberbullying victims. Section 5.3 discusses the study design, while Section 5.4 reports on the recruitment of the co-designers. The design and prototyping activities are detailed in Section 5.5, and Section 5.6 reflects on

the PD approach used in developing the prototype. The limitations of the approach used to develop the application's prototype are considered in Section 5.7. Finally, the chapter's concludes with a summary in Section 5.8.

5.2 Participatory Design

Participatory Design (PD) refers to a set of theories and practises aimed at engaging end-users as full participants in activities leading to technological solutions to real-world problems (Muller and Kuhn, 1993; Clemensen *et al.*, 2007). It started life as part of the Scandinavian workplace democracy movement to include employees in the decision-making process for introducing new technologies into the workplace (Floyd, 1993; Gregory, 2003). It recommends that computerisation in the workplace should be providing workers with better tools to do their jobs rather than the isolated automation of work processes without consulting the workers (Schuler and Namioka, 1993). PD is a diverse endeavour drawing on ideas from various fields, including computer science, psychology, engineering, anthropology, and many more. It has been shown to improve technology acceptance in the design of computer systems for sensitive and disempowered users such as medical patients, young children, the elderly, refugees (Waller *et al.*, 2006; Boyd-Graber *et al.*, 2006; Ruland *et al.*, 2008; Hakobyan *et al.*, 2014) and for bringing together a range of diverse views to improve products and services (Gregory, 2003).

In developing technology-based solutions, it is critical that designers prioritise end-users' needs and do not fall into the trap of designing the solution based solely on their own experiences and preferences. PD affords researchers additional opportunities to learn more about the problem domain while designing the solution and has been widely used in developing a variety of technology solutions for young people (Moraveji *et al.*, 2007; Ruland *et al.*, 2008; Hussain, 2010; Frauenberger *et al.*, 2011; Benton *et al.*, 2012; Read *et al.*, 2014; Iversen *et al.*, 2017). Its use in cyberbullying and online abuse detection system is, however, rare.

Ashktorab and Vitak (2016) used a participatory design approach to work with teenagers to design potential cyberbullying prevention solutions. Similarly, Bowler *et al.* (2014) worked with teens and university undergraduates to propose design recommendations for

cyberbullying intervention for social media sites. Qonitatulhaq *et al.* (2019) utilised a PD approach to collaboratively create an educational video about the dangers of cyberbullying with children aged 9 – 13 years. Finally, McNally *et al.* (2018) employed PD to engage children as co-designers in their study, which proposed design recommendations for parental monitoring software to improve acceptance amongst young people. While the above studies highlight the merits of adopting a PD approach to engaging young people in the design of cyberbullying prevention solutions, in all the studies there appears to be an absence of an actual implementation of the designs and recommendations elicited via PD. There is little evidence to suggest that these few studies that explored the use of PD techniques to ideate cyberbullying prevention solutions actualised these proposals in any form. As such, there is no data on the effectiveness, acceptability or impact of these designs in reality. Furthermore, none of the existing abuse detection tools (Lempa *et al.*, 2015; Vishwamitra *et al.*, 2017; Talukder and Carbanar, 2018; Silva *et al.*, 2018; Oh, 2019) was developed using active end-user participation approaches like UCD and PD. This highlights a disconnect between the active participation of young people in the design of cyberbullying technology to meet their needs and the realisation of such technology for their use and thus empirical evidence of the efficacy of the approach; in so doing, it underscores the novelty of the research reported here in terms of adopting user-centred and participatory design methods to ensure that the needs and desires of cyberbullying victims and young people are heard and taken into consideration when designing and then developing a tool to combat online abuse.

In adopting a PD approach, the aim was to focus the scientific knowledge acquired from literature through the lens of insight gained from various stakeholder engagement activities to ensure that the proposed novel cyberbullying prevention app was designed and developed to address critical areas for the target audience.

5.3 Study Design

FIGURE 5.1 illustrates the design process for the proposed mobile application and highlights the importance of the PD phase to the entire process by serving as the means through which the user requirements gathered via the focus groups and interviews are married to the technical activities required to develop the mobile application.

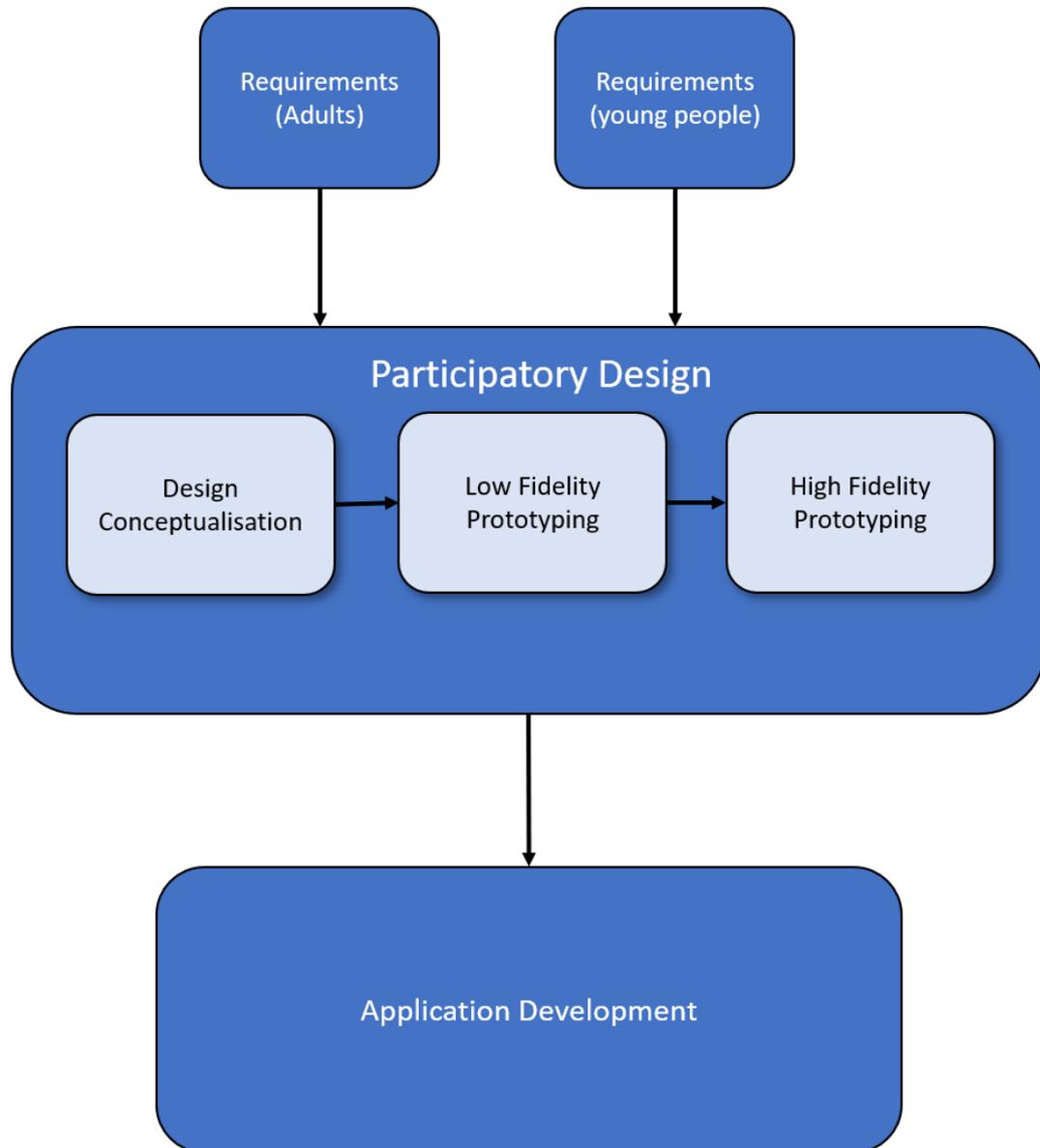


FIGURE 5.1: Overview of the design process.

The PD phase comprises three key stages, namely design conceptualisation, low-level prototyping and high-level prototyping. A prototype is a model created to develop and test ideas. Walker *et al.* (2002) defined fidelity as how distinguishable design artefacts are from the final product and the corresponding ease with which they can be manipulated during the design process. The closer (in appearance) to the final product, the higher the fidelity of the prototype. Low-fidelity artefacts differ from the final product in many ways, including the level of detail, appearance, and form of interaction; their power is that they allow designers to focus on users' interactions with the system instead of visual details (Landay and Myers, 2001) and are very useful for quickly facilitating the expression of ideas and concepts; furthermore, users tend to be more honest regarding their opinions

of low-fidelity prototypes because they appreciate the ease with which changes can be made. A common form of low-fidelity prototype is a paper prototype, created using office tools such as paper, coloured pens, Post-It Notes and markers.

The design conceptualisation stage aimed to devise a high-level logical structure for the application, including screen navigation within the app and how key features can be accessed. The low-fidelity prototyping stage, as described above, made use of simple tools like paper, marker and Post-It notes to add more details to the concepts created in the conceptualisation stage. Its aim was to create a wireframe prototype that provides a general sense of how the app will look like in terms of the 'screens' layout. The high-fidelity prototyping stage creates the final design before the development of the mobile application. It is aimed at producing a prototype that closely resembles the finished product, and that can be used to assess the possible experience of end-users. Three to four sessions were initially planned for the PD phase. The first and last sessions would be dedicated to design conceptualisation and high fidelity prototyping, respectively, and the middle session(s) would focus on low fidelity prototyping. Unfortunately, only two of these sessions could take place. This was due to the difficulty experienced in getting the student co-designers to attend subsequent sessions after the initial session. On a number of occasions, PD sessions were scheduled with the students, and none showed up. Rather than cause further delays to the research program (in addition to those caused by the non-engagement of the secondary school partnered with for the students' interviews – as discussed in Chapter 4), a high-fidelity prototype was developed by the researcher using the outputs of the first session (discussed in Sections 5.5.1 - 5.5.3). Simultaneously, continued attempts were made to engage the students, which eventually yielded a second and final meeting with all six co-designers present. This session was therefore used to review and amend the high-fidelity prototype created.

5.4 Recruitment of Participants

Six of the twenty-five undergraduates interviewed during the previous knowledge elicitation phase (see Chapter 4) were invited to participate in the PD study. Participatory design requires a high level of commitment and engagement from participants; therefore, it is important to select participants to promote active and complementary participation of all members while also ensuring adequate representation. The selected participants were

Id	Gender	Cyberbullying Experience				Ethnicity
		Victim	Bully	Witness	None	
P1	Female				✓	Caucasian
P2	Trans Female	✓		✓		Caucasian
P3	Male		✓	✓		Black
P4	Non-Binary			✓		Caucasian
P5	Female	✓		✓		Asian
P6	Male	✓		✓		Caucasian

TABLE 5.1: Summary of PD participants profiles.

chosen based on observations (during the previous phase) of their enthusiasm as well as their personal experiences of cyberbullying. Additionally, the group's composition was designed such that the key cyberbullying roles (based on their past experiences) were represented. A summary of the participants' profiles is presented in TABLE 5.1

Participants attended two design sessions, each of which lasted up to three hours; each participant was remunerated by means of a £20 Amazon voucher per session. To familiarise participants with the concept and ethos of participatory design, links to youtube videos (see Appendix C.1) on participatory design were emailed to the participants to watch before the sessions. The sessions took place in one of the University's meetings rooms, where drinks and refreshments were provided for all participants. Design tools in the form of standard office supplies (e.g., Post-It notes, paper, and markers) were made available in the meeting room along with a whiteboard and a shared work surface comprising brown paper taped to tables. The meeting room was arranged to have the work surface in the centre of the room, providing ample space for all group members to work together and breakout areas if required at the room's sides (see FIGURE 5.2). A video camera was positioned to capture the work surface, and a separate voice recorder was placed in the room to capture discussions. In addition, the researcher took notes based on the discussions that took place in the sessions.

5.5 Design and Prototyping

As previously mentioned, two participatory design sessions were held with the six participants. The first session was devoted to design conceptualisation and low-fidelity prototyping, while the second design meeting focused on high-fidelity prototyping using

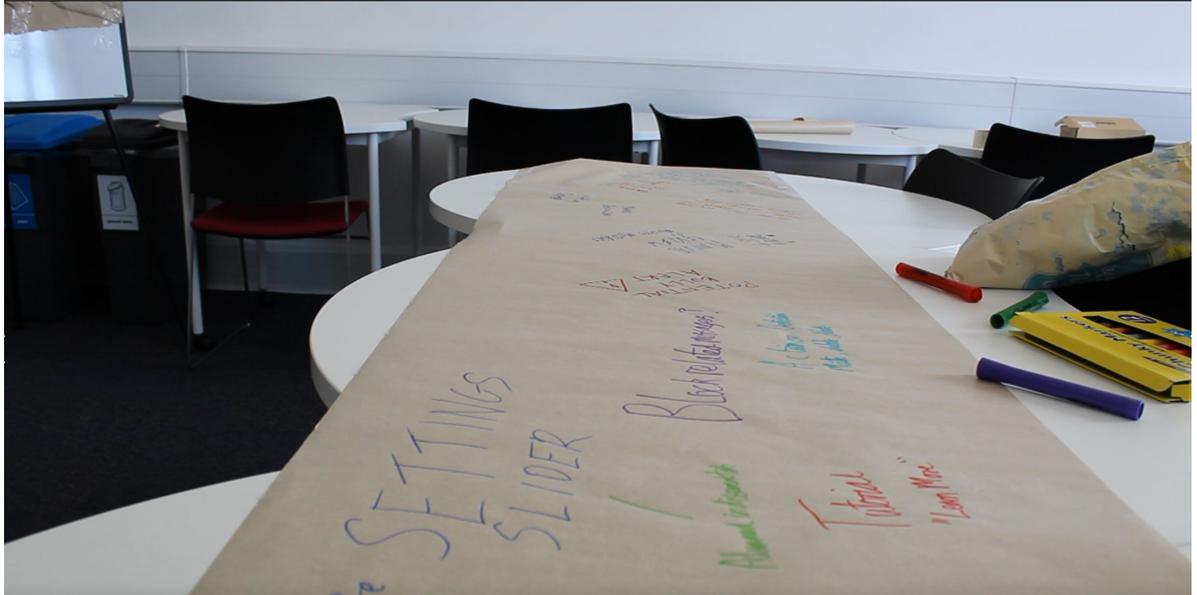


FIGURE 5.2: Room arrangement showing work surface and breakout areas.

proto.io, a mobile app prototyping software. These activities are discussed in detail in the following subsections.

5.5.1 Design Conceptualisation

The first PD session was used to organise the concepts uncovered in the earlier qualitative studies into a single visual representation of the proposed app. The session started with the researcher providing an overview of participatory design, the overall aims of the study, and the specific goals of the design session. This was followed by a brainstorming session where the group collaboratively developed different concepts and ideas for the app and wrote these down on the work surface (see FIGURE 5.3).

Participants were then provided with copies of the combined features list from the focus groups and interviews (see TABLE 4.7) to reflect on. The group reviewed the list, refining, enhancing and removing features. Participants then added items from the features list to the work surface, and a dot voting process was used to prioritise the ideas written down on the work surface. Dot voting is a simple decision-making and prioritising technique used to decide an order amongst multiple items by a group of people. Each group member was assigned six votes and instructed to place a dot next to the item they wished to prioritise. Team members could use as many of their votes as they wished on individual items. This was an interesting and enlightening exercise as it helped uncover



FIGURE 5.3: Participants working collaboratively and writing down ideas.

the criticality of specific features to the participants. It was interesting to discover that, when restricted to a fixed number of votes, some features that participants were initially passionate about received fewer votes than others. For example, whitelisting contacts (thereby instructing the app not to analyse messages received from them) was considered a critical feature by five participants during brainstorming/discussion, yet it received only two votes during the prioritisation process. Similarly, an *on/off* switch that allows users to enable or disable the app without having to uninstall the application and a feature to selectively disable protection for specified social media accounts, which were favourites for many participants, received two and one votes respectively during the process. The features and their associated votes are presented in TABLE 5.2.

After the prioritisation exercise, the group created a spider map of the application using the prioritised features (see FIGURE 5.4). The spider diagram identified vital elements of the app, their relationships, and the high-level navigation between the application screens. The spider diagram revealed additional logical relationships between the app's components necessitating a second review and re-prioritisation of the app's features by the group.

The re-prioritised features list is depicted in TABLE 5.3 (with the old and new rankings) and FIGURE 5.5. This reprioritisation promoted a number of features with low votes. For example, the accounts toggle feature got elevated to the first rank from tenth. The group identified this feature as a unique feature for the proposed mobile app. They believed that

Rank	Feature	Votes
1	Give users the ability to review deleted messages.	5
2	Reassure users that the app is secure and that their personal data is safe.	4
3	Give users the ability to adjust the sensitivity of the offensive message detection	4
4	Provide a friendly and welcoming interface.	3
5	Give users the ability to block contacts manually.	3
6	Allow users to whitelist contacts.	2
7	Provide a toggle switch so that the app can be disabled without uninstalling.	2
8	Provide details of support helplines that cyberbullying victims can call.	2
9	Provide an account toggle switch to selectively enable/disable protection for individual social accounts.	1
10	App tutorial.	1
11	Cool character/logo to represent the app.	0
12	Minimise patronisation	0

TABLE 5.2: Features prioritisation using dot voting.

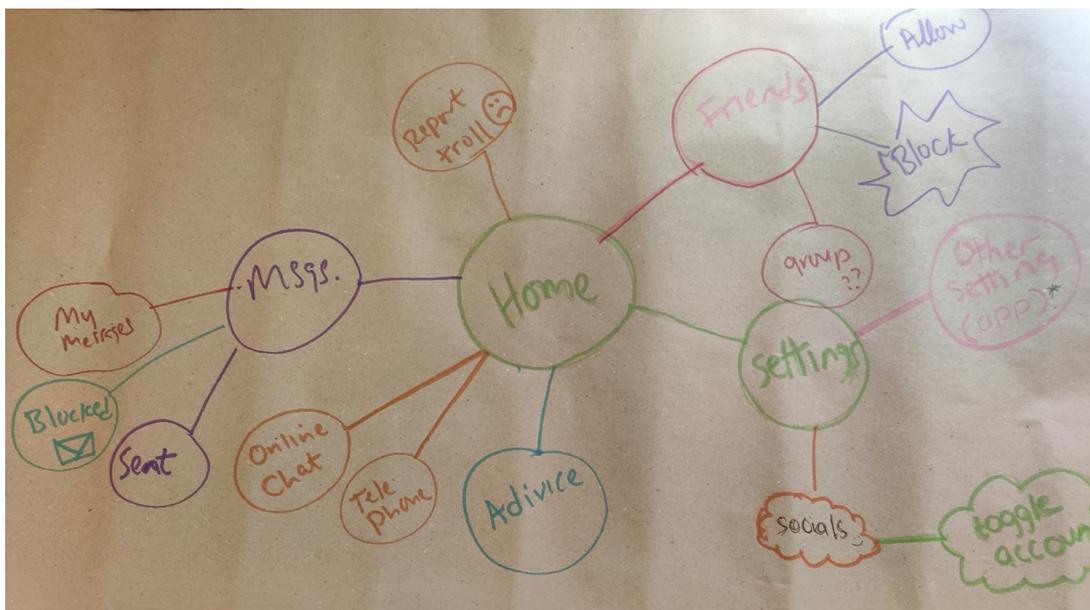


FIGURE 5.4: Spider diagram of key features created during the first PD session.

the feature will endear the app to prospective users as it will allow them to manage the protection of their social accounts more efficiently. It was also noted that the feature would likely reduce the likelihood of users uninstalling the app as they can simply disable and enable the app depending on their online social activities at specific instances. It can also be seen that the contacts whitelisting feature that was a favourite amongst the participants but received low votes was re-instated as a key feature. The ability to adjust the app's

sensitivity to offensive content and a friendly and welcoming UI maintained their ranks while the importance of branding in the form of the app's logo was reinforced by promoting "having a cool logo" from twelfth position to fifth. Three new features were also introduced. These were a setting to prevent the app from sharing details of deleted messages with parents or relevant authorities, not being patronising to the young target audience and the need to reassure them that the app is secure and that their data is safe.

Rank		Feature
Old	New	
10	1	Provide an account toggle switch to selectively enable/disable protection for individual social accounts
7	2	Allow users to whitelist contacts
3	3	Give users the ability to adjust the sensitivity of the offensive message detection
4	4	Provide a friendly and welcoming interface
12	5	Cool character/logo to represent the app
-	6	Provide an option not to send deleted messages to parents or the relevant authorities
-	7	The app should not be patronising of the target audience
5	8	Automatically flag and block abusive contacts.
-	9	Reassure users that their data is safe and that no one will know that they are using the app.

TABLE 5.3: Re-prioritised features.

5.5.2 Low Fidelity Prototyping

After the design conceptualisation activities, the rest of the first session was devoted to low fidelity prototyping. Using the spider diagram as a guide, the group started by creating the

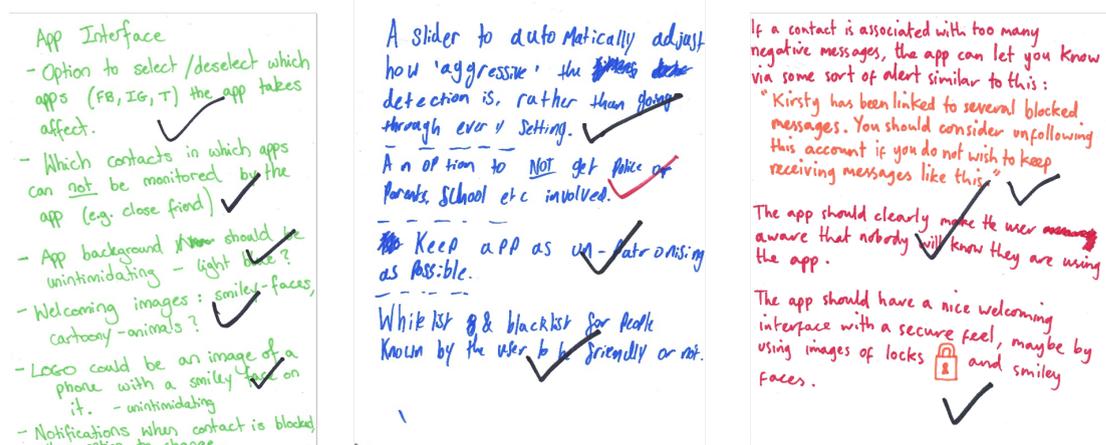


FIGURE 5.5: Essential features shortlist compiled during the first PD session.

prototype for the Home screen and worked outwards to the other screens implementing the prioritised app features. Due to time restrictions, only the Home and Deleted Messages screens could be designed on the day, and the creation of the remaining screen prototypes was planned for following sessions (which did not take place for the reasons discussed in Section 5.3). The participants identified the proposed mobile app's home screen as a critical factor in convincing users to persist with the app on first use. One of them said:

"I've downloaded apps before that when I opened them; I just didn't like the look, what I saw wasn't what I expected. I deleted them straight off."

After discussing ideas for the home screen, some of which were conflicting (for example, a participant wanted to include a scrolling newsfeed-type feature of deleted messages on the home screen while others argued that presenting users with a selection of their deleted offensive messages from the onset might be distressing for users), the researcher proposed the use of Pair Design (Bellini *et al.*, 2005) to facilitate a cohesive start to the home screen design process. Pair Design is a technique used within the digital creative industry to develop design prototypes. It is a method borrowed from the Pair Programming practice (Williams and Kessler, 2003) in software engineering. Pair programming is a software engineering procedure where two software engineers work side-by-side on one computer collaborating on the same programming problem. For several years now, it has been successfully used to deliver higher quality code faster (Williams *et al.*, 2000; Williams and Kessler, 2003). Such was the success of pair programming that it had been adopted by the design industry as a means of delivering higher-quality designs at a quicker pace (Yao, 2015). Pair design is, therefore, by extension, the practice of two designers working together on the same design problem to achieve an improved output. It is sometimes used as a technique to channel the different design perspectives of multiple designers into the creation of competing prototypes of equal importance that can then be deliberated upon (Chapman, 2018). The use of this technique in prototyping the home screen was therefore based on the idea that it would be more productive for three pairs of designers to deliberate on three completed prototypes than for six designers to reflect on the multiple UI elements of a single screen. The researcher has several years of industry experience in pair programming and pair design and facilitated the process. The six participants were paired into three paired groups (see TABLE 5.4), and each group created a version of the home screen. The groups were paired to reflect a combination of competing and complementary views in each pair based on the researcher's observation of the home

screen design's initial discussions. The three designs created from these paired design sessions are shown in Figure 5.6.

Group	Designer 1	Designer 2
Team 1	Participant 1	Participant 3
Team 2	Participant 2	Participant 5
Team 3	Participant 6	Participant 4

TABLE 5.4: Pair Design teams' composition.

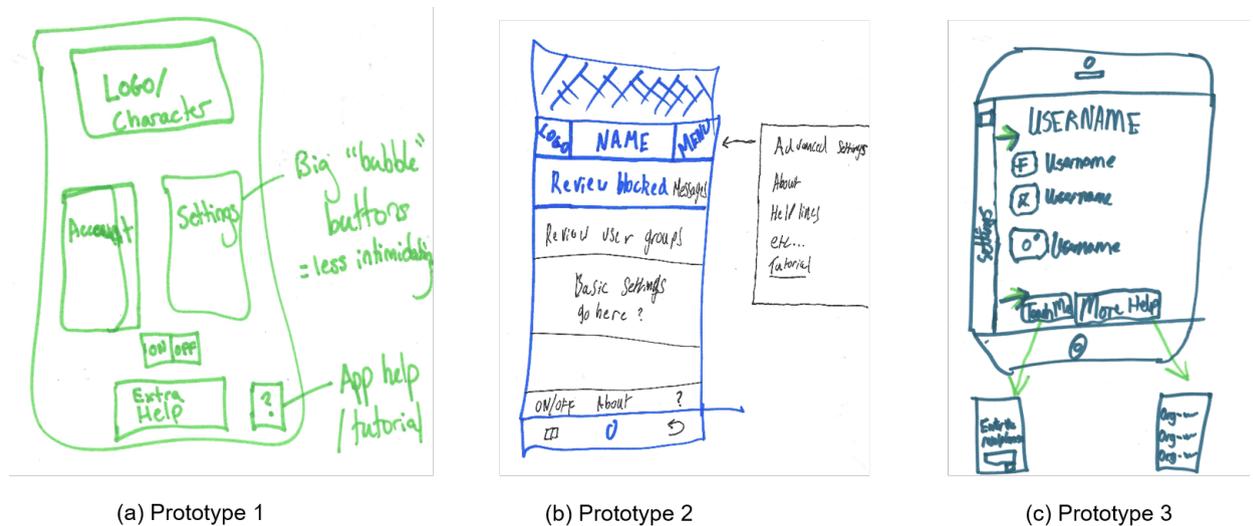


FIGURE 5.6: Home screen prototypes from created using Pair Design.

As the session aimed to create a single design prototype for the home screen, Layered Elaboration (Walsh, 2009) was used to generate a single unified model from the three prototypes. Layered Elaboration is a design technique that enables design groups to expand on ideas presented by others by layering new ideas on top of earlier ones (Walsh *et al.*, 2010). The technique can be likened to adding layers of transparencies with drawings on top of each other. Layered Elaboration has its roots in storyboarding for interactive media and was originally conceived during the design of a motion-controlled video game for the Nintendo Wii video game system (Walsh, 2009).

A base prototype was required to start the Layered Elaboration process. The group reviewed the three prototypes to decide on the prototype to use as the base prototype. After some deliberation, the group settled on Prototype 1 because it shared UI elements with the other two prototypes such as featuring a means to access the help screens (similar to Prototype 3) and the 'On/Off' toggle (shared with Prototype 2). Elements of the prototype that were not universally agreed were removed, and the base prototype was redrawn, as shown in FIGURE 5.7.



FIGURE 5.7: Home screen base prototype.

The design team then took turns to add elements from their prototypes onto the base model, as illustrated in FIGURE 5.8. After some deliberation, Team 2 added four key elements from their design to the home screen's base prototype: buttons to access the screens to manage deleted messages, contact groups, social media accounts and the app's tour. The last team introduced an extended side navigation menu, a button to access the settings screen and smiley faces to indicate "welcoming designs and images" on the home screen.

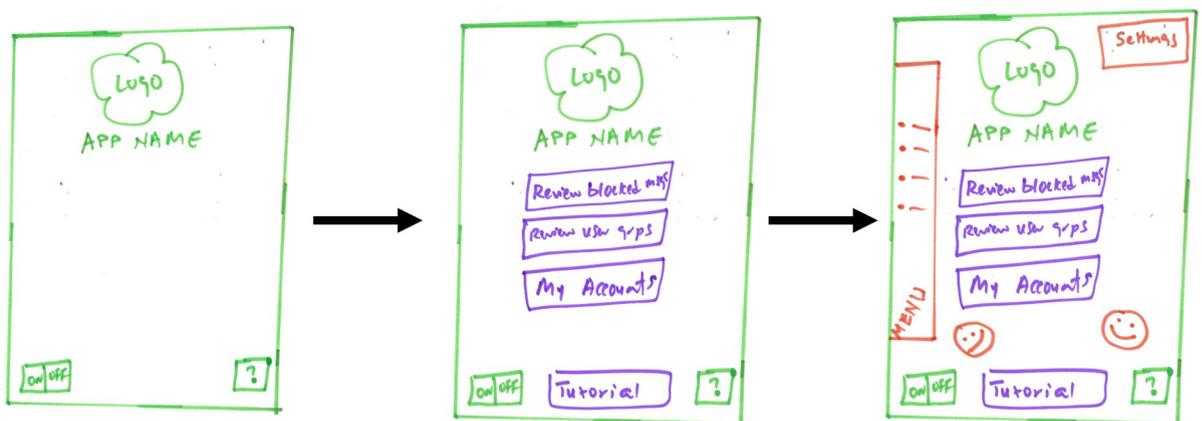


FIGURE 5.8: Home screen base prototype.

After finalising the home screen design, the group created the prototype for the Deleted Messages screen. The impact of the pair design and layered elaboration activities

become immediately apparent as the group's collaboration was more natural and the Deleted Message screen prototype (see FIGURE 5.9) evolved at a faster pace. Due to the similarity between the Deleted Messages screen and the Sent and Received Messages screen, it was suggested that the Deleted Messages prototype should also be used as the basis of the design of the Sent and Received Messages screen.

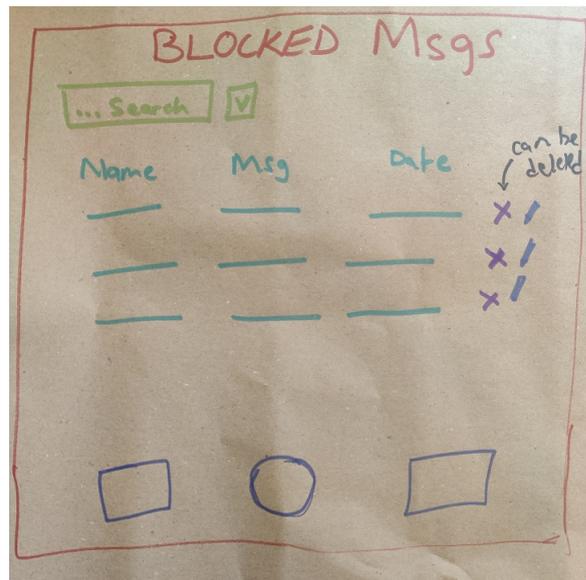


FIGURE 5.9: Deleted Messages screen prototype.

5.5.3 Findings from the First Participatory Design Session

In addition to the design outputs, the voice recording of the group's discussions provided additional data in the form of views, observations and design recommendations. These were transcribed, coded and analysed to discover key themes to guide the high-fidelity prototype design.

5.5.3.1 Use a Cool and Symbolic Logo

A common theme that was repeated by the participants was the need to have an app logo that personifies the app's brand and one that the young target audience can identify with as exemplified by the following comments:

"You should design a cool logo for the app".

"I think especially for like younger children. A nice logo, even like a character, will be attractive to them".

“An image like a padlock to show that the app is safe. Or like a shield. I think my antivirus logo is a shield and people automatically associate that with security”.

These suggestions were grouped under the theme “having a cool logo” and subsequently identified as a key design consideration for the next design phase.

5.5.3.2 The App’s Name Should be Short and Catchy

In addition to the logo, the participants advised that the app’s name should also be symbolic of the app’s core function and complement the logo. One co-designer noted:

“The [app’s] name should be short and catchy. Like Reddit and Pinterest. The name tells you what the app is all about”.

Another participant, however, disagreed with this view:

“I don’t think the actual name matters too much. It can even be silly. The name just needs to be cool. There are so many apps out there that are really popular with stupid names”.

This opinion was supported by another participant, who said:

“Do you remember the Yo app. It was a very stupid app, and all it did was send yo to people. How stupid is that but it became trendy for a short while. So anything can be”.

While the co-designers differed in their opinions on how the app’s name should be derived, there was a consensus that whatever name chosen for the app should be selected to create a strong initial impression with the target audience.

5.5.3.3 Use a Neutral but Friendly and Welcoming Colour like Blue

Some participants suggested the colour blue as a good choice for the app’s interface. In supporting this recommendation, they said:

“Facebook and Twitter both use blue, and I see blue in a lot of apps too”.

“It’s [blue] kind of neutral in a way, so it’s safe. I know it’s meant to attract like kids as well, but if it’s like too bright, yellow and stuff like that, it can be very dividing. Some people will like it and others won’t. Blue is safe, I think.”

5.5.3.4 The App Should Not Look Childish or be Patronising to the Young Target Audience

The co-designers were unequivocal in their recommendations that the app should not be patronising to the target audience. Some of them said:

“Even when I was younger like 12 and 13, I hated anything that we were shown in school that was childish. I remember then all the PowerPoint we were shown were all cartoony with that squiggly text font.”

“That’s quite important. You need to treat them like they are not kids or else they will just say this is another thing that thinks they know us”.

“I remember that at that age, I was quite rebellious. I just didn’t want to listen to adults. Even though I was getting bullied, I didn’t tell anyone. I didn’t read any of the bullying advice then because they looked childish. Cartoon pictures and all that”.

This finding was, therefore, identified as a key design consideration when creating the high-fidelity prototype.

5.5.4 High-Fidelity Prototyping

As previously mentioned, a high-fidelity prototype provides a depiction of the final product that is much closer in visual appearance/form factor to the final product than a low-fidelity prototype. With the use of specialist prototyping software, high-fidelity design prototypes that provide a detailed and accurate representation of the final product can be created quickly. As mentioned in Section 5.3, the proposed sessions to create low-fidelity prototypes for the other app’s screen could not be held due to non-engagement of the students recruited as co-designers. Using the spider diagram and the Home and Deleted Messages screen prototypes produced in the first session as a guide, the researcher used the web-based prototyping tool, proto.io, to create an interactive prototype of the app. Proto.io is web-based software that allows people with little to no design and software programming knowledge to create detailed interactive prototypes for mobile applications. It provides a simple drag and drop interface that allows designers to quickly create mobile app prototypes that can be shared and viewed on mobile devices. The spider diagram was subsequently refined by the researcher during the high-fidelity prototyping phase to accommodate the practical re-arrangement of some screens during the prototype creation (see Figure 5.10).

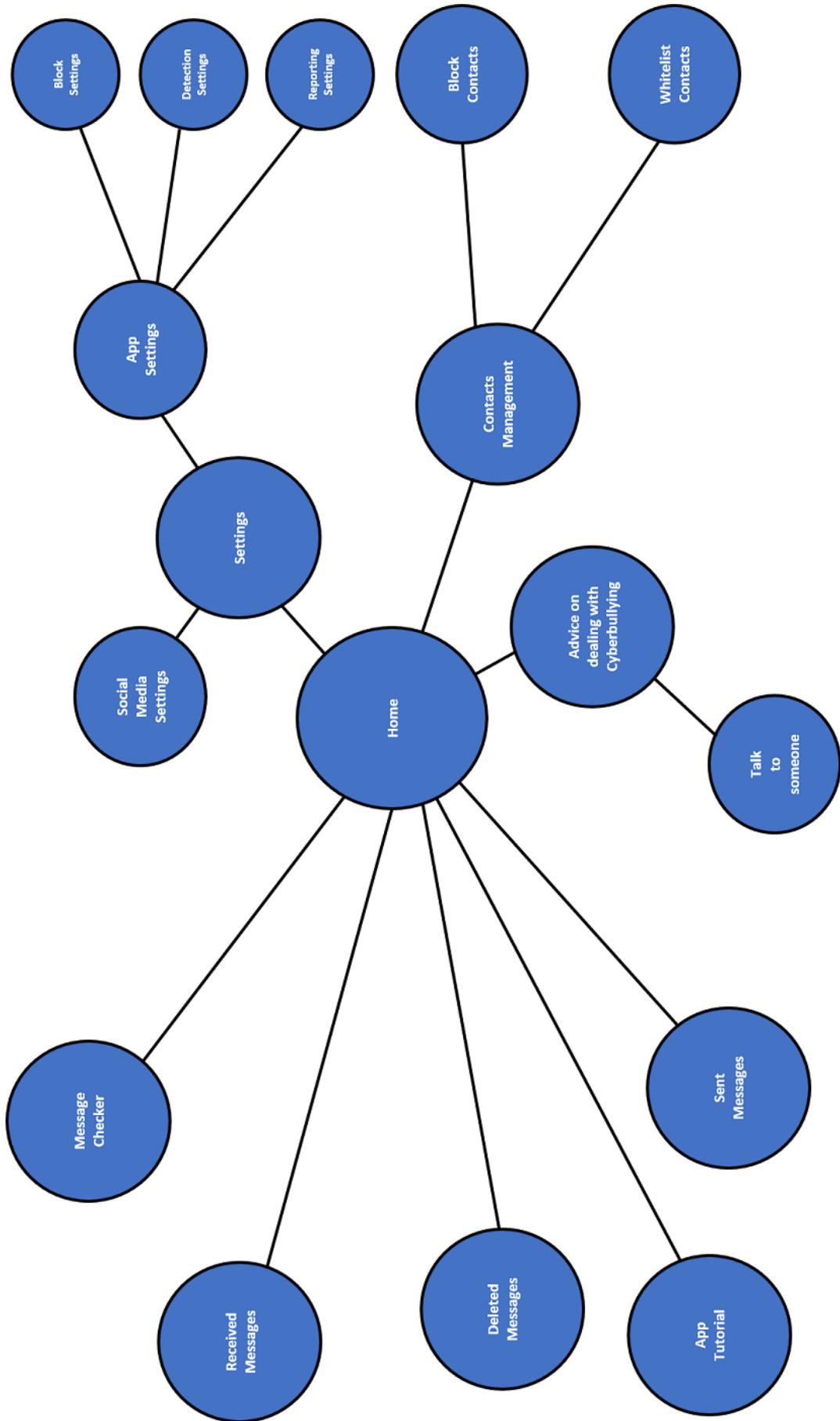


FIGURE 5.10: Spider diagram of key features for the proposed mobile apps.

As the Deleted Messages screen shared many features with the Sent and Received Messages screen, the low-fidelity prototype created in the first session was used as the foundation of all the app's messages-related screens. Additionally, the findings and co-designers' design suggestions from the first session served as input into the interactive high-fidelity prototype created by the researcher: in essence, in lieu of a full participatory design approach, an evidence-based user-centred design (UCD) approach was used to ensure the progression of the research agenda. Ultimately, the second design session was, therefore, aimed at reviewing and refining this interactive prototype with the co-designers.

The session began with a review of the previous session, the study aims, and the session's objectives. The group used two Windows laptops and one MacBook to work on the interactive prototype. Another advantage of using the proto.io tool is the ability to view changes made in real-time via a companion mobile app, the co-designers therefore downloaded and installed the proto.io mobile app onto their phones. This allowed them to view the changes made on the laptops on their mobile phones in real-time, providing them with a mobile-based experience of the prototype similar to the way prospective users will experience the proposed mobile app. The researcher presented a walkthrough of the prototype using the updated spider diagram and starting from the first screen displayed after the app was installed. The group collectively reviewed and amended each screen as they progressed through the app. In addition to individual screens, the overall design of the app's interface was reviewed as well, especially in terms of the group's design suggestion from the first session. The group responded positively to the app's use of a stylised shield as a logo (see FIGURE 5.11) and the app's name (BullStop).

They believe that the app's name and logo would help it stand out in the app store and attract the attention of the intended audience. The co-designers identified with the overall design of the prototype's user interface and said it was not patronising or childish. They welcomed the use of the colour blue in the prototype, as suggested in the first session and commented that the colour makes the prototype feel familiar like social media apps they have used before. While the PD sessions supported the ideation/generation of a high-fidelity prototype that reflects the target audience's vision for the app, further refinements were still required to ensure that the embedded UI elements conform to sound HCI principles.



FIGURE 5.11: The logo for the Proposed mobile app.

The Splash screen (FIGURE 5.12) is designed to be presented to users the first time the application is launched as the mobile phone loads the app into its memory. As most phones' computing capabilities are such that apps are loaded into memory quickly, this process should be unnoticeable to most users. The EULA (End User License Agreement) (FIGURE 5.13) will therefore be the more likely screen that most users will see after installation. The user will then be prompted to read and accept the license agreement (this is required to use the app) as well as the app's privacy policy (see FIGURE 5.14).



FIGURE 5.12: Splash Prototype.



FIGURE 5.13: EULA Prototype.



FIGURE 5.14: Privacy Prototype.

On accepting the license agreement, the user can create a BullStop account (FIGURE 5.15). The account will be used to store the user's profile and application settings. Once the account has been successfully created, the user will be presented with the home screen (FIGURE 5.16) from where other areas of the app can be accessed including the app's settings (FIGURE 5.17) which can be accessed via the cogwheel in the top right corner of the screen.

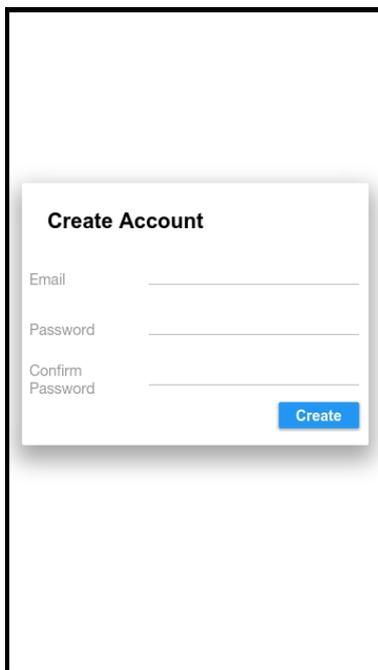


FIGURE 5.15:
Account Prototype



FIGURE 5.16:
Home Prototype.

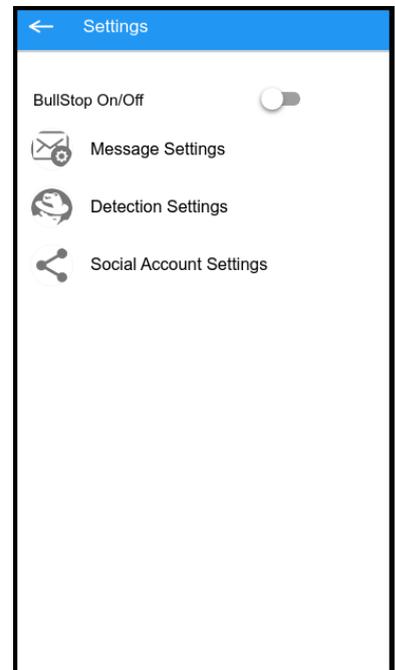


FIGURE 5.17:
Setting Prototype.

From the Settings screen, the user will be able to access additional settings such as the Message (FIGURE 5.18), Detection (FIGURE 5.19) and Social Account (FIGURE 5.20) Settings screens. From these settings screens, the user would configure the app's operations including how deleted offensive messages are handled, the app's sensitivity to objectionable messages, how abusive users are handled and connecting social media accounts to the app for protection.

Figures 5.21, 5.22, 5.23 depict the Received, Sent and Deleted Messages screen prototypes, respectively. The Received and Sent Message screens would list messages synchronised from the social media accounts while offensive messages automatically deleted by the app will be accessible via the Deleted Messages screen.

The Message Review window (FIGURE 5.24) will be displayed as a pop-up screen when the user taps on any message in the messages-related screens and will allow users to



FIGURE 5.18:
Message Settings
Prototype.

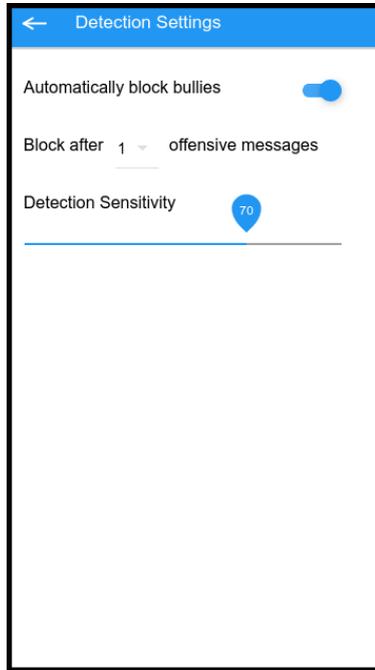


FIGURE 5.19:
Detection Settings
Prototype.

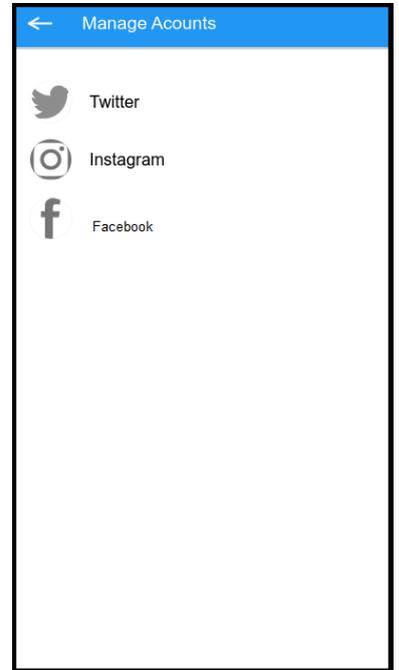


FIGURE 5.20:
Social Account
Settings Prototype.



FIGURE 5.21:
Received Prototype.



FIGURE 5.22: Sent
Prototype.

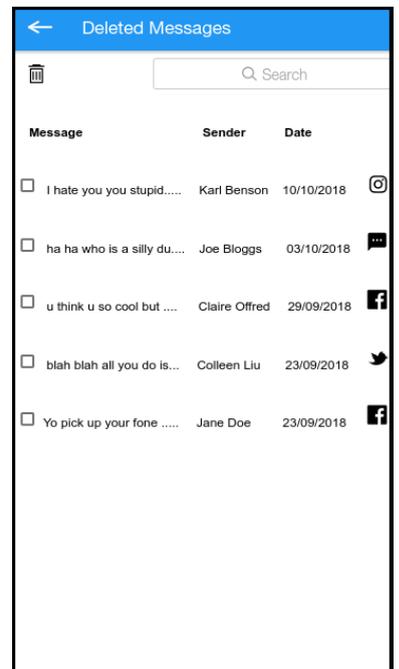


FIGURE 5.23:
Deleted Prototype.

review and update the offensive labels assigned to a message. The prototype of the Message Checker, a reflective tool designed to help users determine the appropriateness of messages they are about to send, is shown in FIGURE 5.25.

The app will allow users to manage contacts retrieved from their social media accounts through the Contacts screen (FIGURE 5.26) and the Help screen (FIGURE 5.27) will allow users to access the application’s tutorial pages (FIGURE 5.28) and the anti-bullying charity organisations’ helplines (FIGURE 5.29).

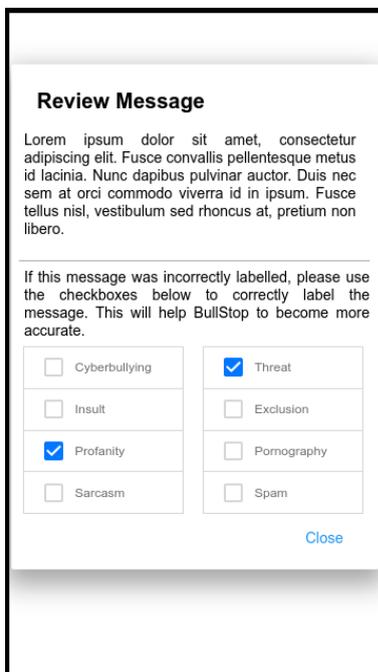


FIGURE 5.24:
Message Review
Prototype.

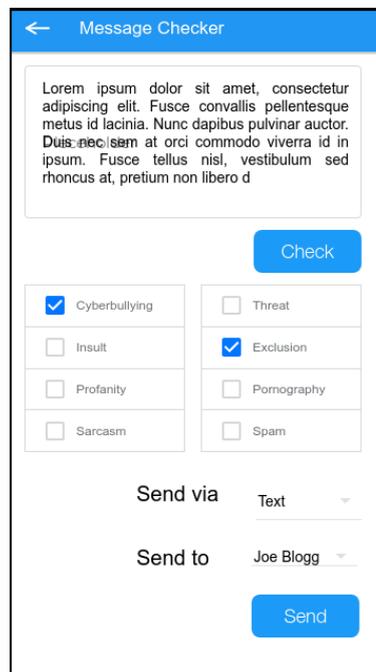


FIGURE 5.25:
Message Checker
Prototype.

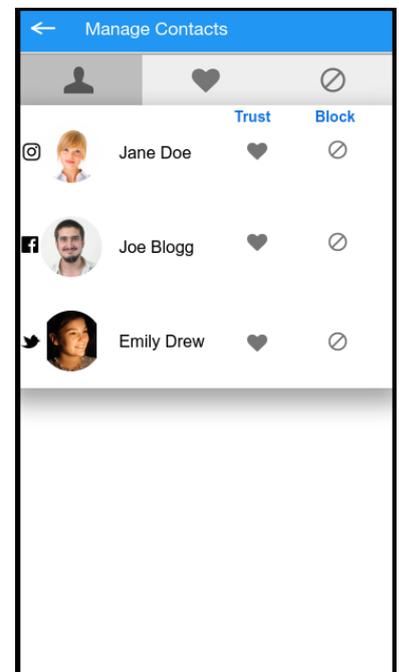


FIGURE 5.26:
Manage Contacts
Prototype.

Finally, the prototype of a sample error message is shown in FIGURE 5.30.



FIGURE 5.27: Help Prototype.

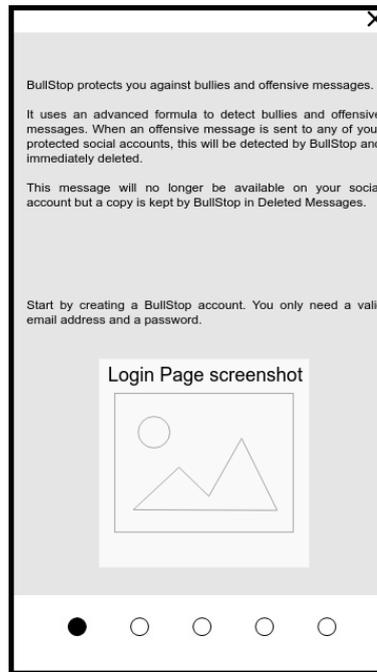


FIGURE 5.28: Tour Prototype.



FIGURE 5.29: Helplines Prototype.

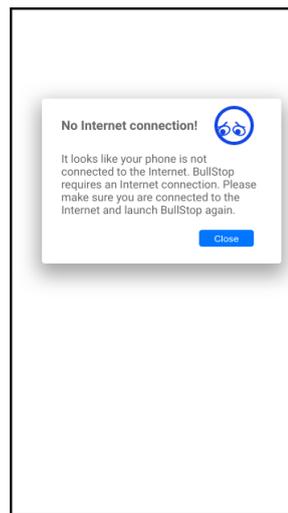


FIGURE 5.30: Error Message Prototype

5.5.5 Findings from the Second Participatory Design Session

The analysis of the second PD session transcripts unearthed further design considerations to guide the final stage of development of the proposed mobile application.

5.5.5.1 Provide Shortcuts to Key Components to Facilitate First-time Use

After interacting with the prototype for some minutes, one of the co-designers made the following observation:

“I tapped on the disabled text ‘cause it was in red. I expected that to do something, but nothing happened”.

The group reflected on this and agreed that it is natural for users to expect that tapping on the ‘Disabled’ text on the home screen should perform an action (in this case, enabling the app so that social media accounts can be protected). The students also recommended that the social media icons at the bottom of the home screen should serve as shortcuts to link the relevant social media account to the app. They reasoned that the inclusion of these shortcuts would accelerate the app’s initial set-up for first-time users. These UI elements are indicated (in red outline) in FIGURE 5.31.

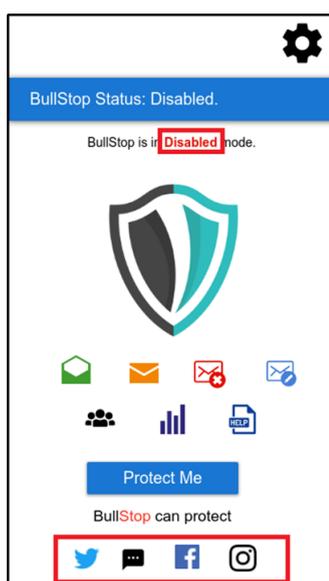


FIGURE 5.31: Suggested shortcuts

5.5.5.2 Use familiar Icons to Signpost Actions

While the co-designers did not experience issues identifying the functions associated with the app icons, a few commented that the Received and Sent Messages icons might confuse some users. They suggested exploring alternative icons that would be universally recognisable to represent these functions.

5.5.5.3 Reassure Users that the App is Secure

“You’ll need to convince people that their socials are safe with your app”.

The above comment highlighted a key challenge that the proposed app may encounter – that is, convincing users that the app is secure and that their data is safe when using the app. Another student noted:

“Probably the worst thing that can happen to you online is to get your account hacked, and I think that would be one of the first things on people’s mind. Can my account get hacked through this app?”

One co-designer, however, had a different view:

“To be honest, I think any account can get hacked. I mean Twitter accounts get hacked all the time, so I’m not sure what extra stuff the app can do”.

It was therefore decided that the app should, as a minimum, implement standard online security measures including:

- (i) enforcing the use of strong passwords;
- (ii) verification of identity via email address; and
- (iii) allowing password reset only via email.

5.5.5.4 Emphasise that the App Supports Multiple Social Media Platforms

The participants believed that integrating with multiple online social networks is a critical proposition of the mobile app and one that should be emphasised.

“When I saw the friends page, I thought that’s cool. To have all your friends from your socials in one place is quite nice”.

A similar comment was made about the messages screens:

“People can get bullied on different platforms. My ex and her friends sent me Snapchats and posted comments on my insta. So if everything is deleted and in one place, it’s easier to handle”.

When challenged to propose ways that the app can do this, the group suggested that the social media icons displayed on the home screen should also be emphasised on the app store listing and highlight the various social media platforms supported on the app's page on the store.

5.5.5.5 Creating a Comprehensive Online Presence for the App can Reassure and Attract Potential Users

Expanding on the above suggestion, the group advised that establishing an online presence for the app across the various social media platforms can help reassure potential users of the app's authenticity and assist in attracting potential users. Two of the students noted:

"It's important to have a brand for the app. You can use that to market the app".

"If it [the app] popped up in their feeds then I think they are more likely to trust it".

Overall, there was a strong suggestion from the discussions that reassuring potential users that the app is secure is critical to its acceptability amongst the target audience. This validated similar findings from the earlier work about the need to project a safe and secure image for the proposed mobile app.

5.6 Reflections on Participatory Design with Young People

Designing technological solutions for young people is challenging as the way they interact with technology is often vastly different to adults (Ashktorab and Vitak, 2016). Existing literature on online abuse detection systems for young people displays a conspicuous lack of involvement of young people in any phase of the design process. The study reported here aimed to address this by directly involving young people in the design process and, in so doing, attain a deeper understanding of their specific needs which could then be translated into tangible requirements that could be implemented in the mobile app. The qualitative studies conducted prior to the PD sessions helped establish a set of core features for the mobile app.

In the first PD session, the students were tasked with suggesting key features for the app without seeing the features list compiled based on the earlier studies. This results of this exercise were complementary to the core features list, indicating that the qualitative studies comprehensively captured the features perceived as critical for the mobile app and provided validation that these features are indeed relevant to the target audience. An interesting observation about the two features lists is that, while the features suggested as a result of the earlier studies were in the form of general ideas and concepts (e.g., “Work with multiple social media platforms”), those suggested during the first PD session included implementation details (e.g., “Provide an account toggle switch to selectively enable/disable protection for individual social accounts”). This difference highlights an additional benefit of going beyond focus groups and interviews to actively engage stakeholders in the design process.

In assuming co-designer roles, the participants became more invested in the process and began to visualise the app as they would like to see it implemented. This became more apparent as the session progressed to the dot voting activity. This exercise was extremely useful in identifying critical features for the young stakeholders as, when forced to choose, some of the initially popular features were demoted by the students. By encouraging an in-depth reassessment of the application’s features in this manner, co-designers became more aware of their roles as representatives of young people in the design process and considered features more holistically in terms of their overall usefulness to young people rather than just based on their personal preferences. Interestingly, the features list was reviewed again and the app’s features reprioritised following the spider diagram activity where the group mapped out the logical relationships between the app’s components. The group’s multiple reprioritisations highlight the importance of adopting an iterative review process with stakeholders to identify the critical components of the system necessary for users’ acceptance.

A democratic design approach, such as participatory design, assigns equal rights to all group members and sometimes conflicting visions and personalities can threaten to slow or even derail the design process. The use of pair design, while not a PD method per se, allowed the group to progress past the initial acclimatisation stages as group members collaborated for the first time. Working in pairs allowed the group as a whole to quickly progress from debating six different design perspectives for the home screen to working with a partner to create a shared vision. Alongside pair design, layered elaboration was

instrumental in achieving a unified vision during the design process. It emerged as a design method that is a natural fit for collaborative design. It is hugely surprising that the use of technique did not expand beyond the work of Walsh *et al.* (2010). This is a shame as this study found it extremely useful in facilitating the condensation of the three home screen prototypes into a unified representation of the co-designers visions. Future researchers engaged in similar endeavours would do well to explore the use of this technique.

The merits of adopting a PD approach were apparent in the design sessions, as the interactive nature of the sessions allowed a free flow of creative ideas that may not have necessarily occurred to the researcher if the app was designed without end-user involvement. It also represented a learning experience for not only the co-designers but the researcher as well. While the benefits of the approach were later validated via the evaluation studies (see Chapter 7), co-designers, opinions about their involvement in the process provided some early insight into the value (from their perspective) of the use of PD approaches in context such as this, and these are discussed in the following sections.

5.6.1 Self-pride

The participants said that they felt “honoured” and “special” that their opinions about cyberbullying were solicited and that they have been directly involved in designing a mobile app to aid its prevention. Many were proud that their contributors would be used to help young people experiencing online abuse. One of them noted:

“It feels good to be able to do something for people being abused, especially as I experience it myself”.

Others said:

“I never thought that like you know people are researching stuffs like this [cyberbullying] and it’s really nice to be part of it”.

“I will look forward to the app, and I can tell people [that] I worked on that app”.

Some of the students admitted that the self-pride they felt taking part in the study contributed to their willingness to continue with the study after the first participatory design session:

“I know I was difficult to get hold of for a while [...] I felt like we started something and I was a part of that, and I want to see how the app looks like.”

“I really liked the first session, and from it, I knew a lot more about designing apps and stuff. As I’m studying Computer Science and Biz [Business] I wanted to know more. I felt like this is something I could look into for projects and stuff.”

5.6.2 Learning

The learning aspects of the experience also emerged as a key outcome for the co-designers, as evidenced by some of their comments:

“It was quite interesting to learn how we all came up with different ways of doing the same thing. And I think in the future I’ll try to apply that in like group assignments because usually, I’m like why are you guys so slow, can’t you see what I’m saying but you know, I saw that everyone has a valid point and it’s based on how you look at things”.

“Using the software was quite good. I have never done anything like that before, and it showed me a lot on how apps are created”.

Another student said he wants to learn more about participatory design and write about it for his university coursework. The students were also curious about the researcher’s experience with PD and were surprised to discover that this was the first time the researcher was using the approach. One of the students wondered why the method is not used more often. The researcher cited the difficulty experienced in arranging sessions for the study as a typical example of why developers are wary of using user-centred approaches. The group briefly discussed how such issues could be mitigated, and some of their suggestions are presented in Chapter 8.

5.6.3 Empowerment

Some students said they gained more confidence in their abilities as a result of taking part in the design sessions. One of them said:

“I got a confidence boost from doing the app’s design. I didn’t think it was something I was capable of”.

Another co-designer spoke about previously having ideas for mobile apps but being unsure how to represent these visually:

“I will definitely spend more time thinking about my app ideas and maybe use the proto software to sketch them out”.

5.7 Study Limitations

Part of the novelty of this research program is the engagement of young people as co-designers in the design of the proposed cyberbullying prevention mobile application. The process resulted in a high-fidelity prototype that encapsulated the target audience’s requirements as represented by the co-designers. It was, however, not without its limitations. Common to many PD studies were difficulties experienced in arranging successive meetings with the co-designers to complete the low-fidelity prototype of the proposed app; this necessitated the process of designing many of the high-fidelity prototype screens without the benefit of an equivalent low-fidelity, co-designed model and could be argued as a divergence away from true PD and thus a limitation of the work. While the decision to eschew the low-level models was borne out of pragmatism to prevent further delays to the research programme, it is recognised that there is a risk the high-fidelity prototype may have been different had the co-designers been more fully involved in lower-fidelity prototyping as was initially planned. Additionally, as mentioned in Section 5.3, users may have been less honest regarding their opinions of high-fidelity prototypes due to their perception of how much effort would be required to make changes to the prototype.

Upon reflection, however, the limitations noted above are mitigated by a number of factors. Firstly, all the co-designers’ design suggestions from the first session were implemented in the prototype and subsequently reviewed and validated by the co-designers themselves in the final session. Secondly, the co-designers actively reviewed and refined all the screens of the high-fidelity prototype. During this process, the co-designers actively contributed to the prototype design – adding, enhancing and removing UI elements and reshaping the prototype into an accurate reflection of their vision. Lastly, the researcher demonstrated the ease with which the high-fidelity prototype could be amended during the second session and encouraged the co-designers to use the prototyping tool to amend the prototype as they deemed fit. This was done to ensure

that co-designers did not hesitate to make changes to the high-fidelity prototype and was effective at achieving this goal.

In using a hand-selected sample of Aston University's first-year undergraduates, it could be argued that the sample may not be representative of the larger population of young people. Collaborative design approaches, such as participatory design, require a high level of commitment and engagement from participants, and substantial rapport is necessary amongst co-designers and the researcher for the process to work. The rapport established from the onset during the interviews and the relationship built with the students by the researcher facilitated the success of the PD sessions. Indeed, engagement with participants to build rapport ahead of PD activities was one of the key recommendations of Lumsden et al. (2017) for designing assistive technologies using participatory design. Even with this established rapport, ongoing non-engagement on the part of the co-designers was, as noted above, an issue. It is felt that it would not, therefore, have been possible to conduct the study with an entirely random sample selected from the larger population of young people who had no prior engagement with the research programme.

The final potential limitation of the study relates to the ages of the co-designers, who were older than the *typical* adolescent, and thus their appropriateness as a representative sample for the primary target audience could be questioned. It is felt that the generalisability of the resulting design and the group's suitability in terms of representing the design visions of young people in general have been validated during the evaluation studies (see Chapter 7) and as such it is not felt that this limitation is significant.

5.8 Summary

Understanding the needs of end-users is critical to the successful adoption of a product. The researcher's experience with the use of participatory design as a design technique has been very positive. This view was shared by the participants who found the experience to be empowering and rewarding. By involving adolescents with different experiences of cyberbullying, the study gained multiple perspectives on cyberbullying. This allowed the app's design to be tailored to accentuate features essential to the target audience.

This chapter reports on the participatory design of the cyberbullying prevention application. It details the collaborative design activities conducted with the co-designers leading to the creation of a high-fidelity prototype for the proposed mobile application. PD requires a high level of engagement and rapport amongst participants, and the interviews conducted before the PD sessions helped establish a positive relationship with the co-designers. Additionally, the interviews and focus groups provided an initial set of requirements for the app that was built on and expanded in the PD sessions. The result of the use of PD in this study was a high-fidelity prototype that captured participants' desires for a cyberbullying prevention tool and, by extension, the desires of the broader target audience.

Young people are naturally creative and expressive and, like previous studies (Ruland *et al.*, 2008; Benton *et al.*, 2012; Iversen *et al.*, 2017), PD was found to be a natural fit when designing for this audience. Rather than restricting their creative inclinations, participants were allowed to flourish in the first design session and then collaboratively finalised the high-fidelity prototype in the final session. While young people as the primary audience for the proposed app served as co-designers, the adults participants (especially the mental health professionals) also contributed to the design process. This was in the form of comments, feedback and suggestions provided after the high-fidelity prototype was shared with them and some of these helped shaped the final version of the app's user interface. For example, in an early iteration of the app, the deletion threshold (i.e., the value on which basis a message is deemed offensive and automatically deleted) setting was initially implemented as a drop list with 3 discrete values (i.e., low, medium and high) but following feedback from the psychologist that implementing the deletion threshold setting as a continuous range (rather than 3 discrete values) would be more empowering to online abuse, the deletion threshold setting was re-implemented as a slider control instead of the drop list. The psychologist's suggestion was based on their view that the physical act of increasing (or decreasing) the deletion threshold through the range of values could be an empowering activity for online abuse victims as this is providing them with a high level of control on the type of messages they receive through the configuration of the threshold at which offensive messages are deleted. This highlighted the importance of involving all stakeholders in the design process thereby ensuring that all stakeholders' views are adequately considered and incorporated into the final product.

In the next chapter, the technical development of BullStop – the cyberbullying prevention application developed based on the activities reported in this chapter, is discussed.

Chapter 6: The BullStop Mobile Application

6.1 Introduction

As discussed in Chapter 2, considerable research effort has been invested in the development of algorithms and models to automatically detect cyberbullying and online abuse; there is, however, a conspicuous lack of research focused on the design of systems and tools capable of utilising these algorithms for real-world use. A key objective of this research program was, therefore, the design and development of a computational system for identifying and preventing cyberbullying and online abuse built on findings drawn from academic research and commercial computing tenets. The combination of these two areas resulted in the development of BullStop, a mobile-based cyberbullying prevention system.

This chapter describes the architecture and components of BullStop, the cyberbullying detection and prevention mobile app developed as part of this research programme. It begins, in Section 6.2, with a review of the application requirements, as suggested by the stakeholders, and their implementation status. This is followed by a description of the application's high level architecture in Section 6.3 and an overview of the Android Application Framework that serves as the mobile application's operating system in Section 6.4. Sections 6.5 – 6.8 describe the various application components, while the technical challenges faced in developing the app and the limitations introduced as a consequence are considered in Section 6.9. Finally, Section 6.10 concludes the chapter.

6.2 Application Features Implementation

The UCD and PD activities conducted (as discussed in Chapters 4 and 5) generated a set of application requirements and a high-fidelity prototype of the mobile application. These then served as inputs into the development phase of the research programme. In developing the mobile application, technology constraints and practical considerations enforced compromises in the implementation of some of the application functionalities, resulting in a number of features not being included or implemented in the manner initially envisioned.

One such compromise was the choice of social media platforms that the app can be used with. Twitter is currently the only social media platform supported by the application, and this is due to restrictions on the APIs of Facebook and Instagram; this is discussed in Section 6.8. As a result, the account toggle feature became irrelevant (as there is only one social media account in use) and was subsequently not implemented. Automatically reporting abusive users could also not be implemented because the abuse reporting facility on Twitter (and other social media platforms) cannot be used in an automated fashion (i.e., there is no API for BullStop to call to report a user). Other suggestions that were not implemented in the current version of the app are: displaying daily motivational quotes which was considered a non-prioritised requirement that could be delegated to implementation in a future version of the application; a social media 'safe browsing' mode that was technically impossible to implement; a social media 'time out' mode which required highly intrusive security permissions on the mobile phone to be implemented and was rejected by the Google App store; reporting users to law enforcement which cannot be done automatically and was technically impractical as there are at least 45 different territorial police forces in the UK; and restricting access to websites, which is a parental monitoring feature that is beyond the app's focus.

TABLE 6.1 presents the originally suggested application features, along with an indication of the implementation status of each.

Prioritised	User Requirements	Suggested By		Implemented
		Adult	Young People	
Y	Automatically delete offensive messages		✓	Y
Y	Users should be able to review deleted messages		✓	Y
Y	Keep copies of deleted messages		✓	Y
Y	Reassure users that the app is secure and that their personal data is safe		✓	Y
Y	Give users the ability to adjust the sensitivity of the offensive message detection		✓	Y
Y	Provide a friendly and welcoming interface		✓	Y
Y	Automatically block offensive users	✓	✓	Y
Y	Give users the ability to block contacts manually		✓	Y
Y	Allow 'whitelisting' of contacts		✓	Y
Y	Provide a toggle switch so that the app can be disabled without uninstalling	✓	✓	Y
Y	Provide details of support helplines that cyberbullying victims can call		✓	Y
Y	Provide an account toggle switch to selectively enable/disable protection for individual social accounts		✓	N
Y	App tutorial		✓	Y
Y	Cool character/logo to represent the app		✓	Y
Y	The app's design should not be too childish		✓	Y
N	Report offensive users to the social network	✓	✓	N
N	Display daily motivational quotes	✓	✓	N
N	Provide access to relevant advice and help for cyberbullying victims	✓	✓	Y
N	Permanently remove deleted messages after a configured period		✓	Y
N	Safe browsing mode when using social media apps		✓	N
N	Social media 'time out'		✓	N
N	Online behaviour scorecard		✓	N
N	Report offensive users to law enforcement		✓	N
N	Content filters to restrict access to inappropriate websites		✓	N

Prioritised	User Requirements	Suggested By		Implemented
		Adult	Young People	
N	Work with multiple social media platforms		✓	N

TABLE 6.1: Overview of suggested application features and their implementation status.

6.3 Overview of the BullStop Mobile Application

BullStop comprises two application sub-systems: an android mobile app that serves as the primary interface for users and a cloud backend that houses a number of components, including the BullStop APIs, the machine learning model, and the remote database. The cloud backend is hosted on the Microsoft Azure¹ cloud platform. A logical depiction of the two sub-systems regarding the flow of data is presented in FIGURE 6.1. The arrows in the diagram indicate the movement of data within the system, and it can be seen that data flows from the social media platforms into the cloud backend end via the BullStop APIs or from the mobile app. The data is then sent to the ML model for prediction and the results stored in the remote database, which is synchronised with the mobile app.

FIGURE 6.2 provides a more detailed view of the system, illustrating key components and activities. The mobile app allows users to perform actions such as linking their social media accounts to the app, setting personal preferences, managing messages and contacts and more. A local SQLite² database is used to store the app's data locally on the smartphone. The app connects to online social networks via APIs and regularly (every three minutes) retrieves new and updated data such as messages and contacts from the social media platforms. The retrieved messages are sent to a message queue in the cloud backend, where they are retrieved for analysis by the Abuse Detection Module which contains the ML models used for prediction. Each predicted label (see Section 3.3.3) is associated with a weight, the cumulative value of the predicted labels' weight is the offensiveness score for the message. Users can configure a threshold value for this score and the Marshaller automatically deletes any messages with an offensiveness score equal to or above the threshold by initiating the relevant function via the social media platform APIs. Additionally, the sender of an offensive message can be blocked if so configured by the user.

¹azure.microsoft.com

²sqlite.org

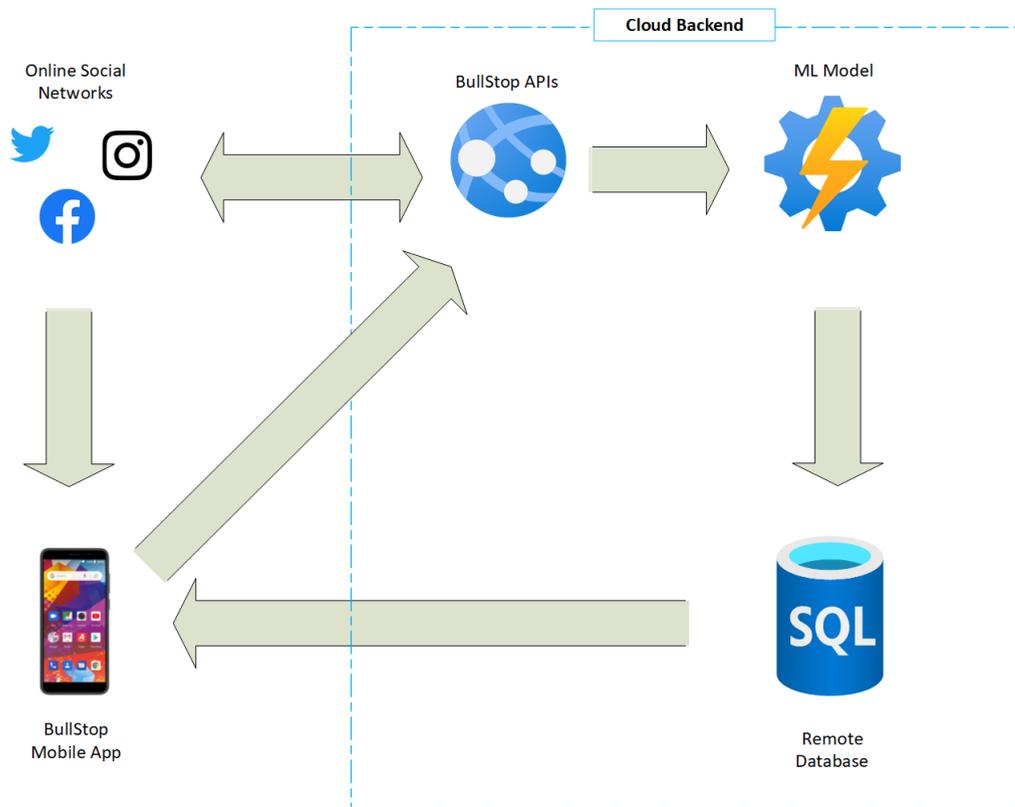


FIGURE 6.1: Logical overview of the BullStop Application System.

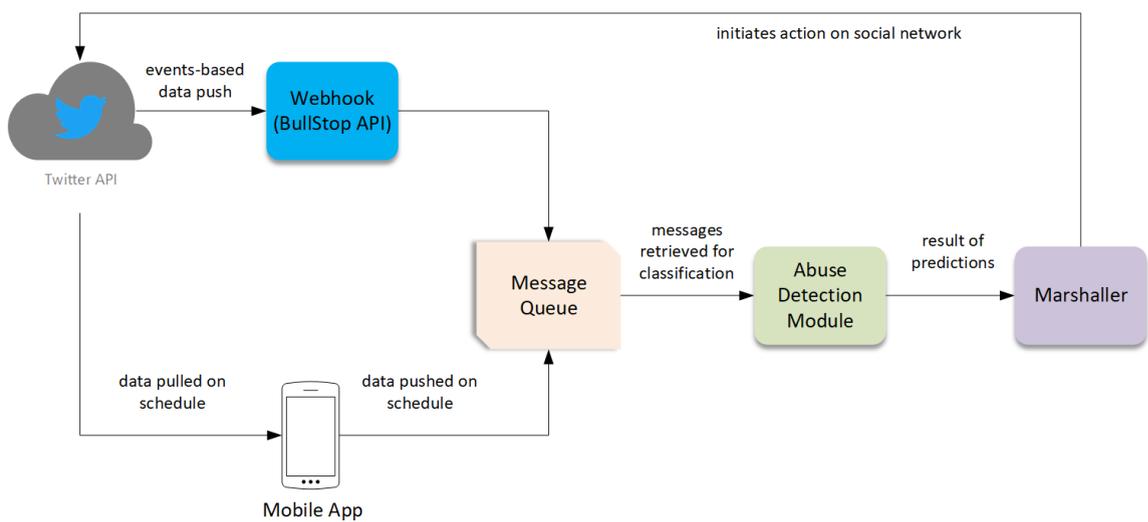


FIGURE 6.2: Key components of the BullStop System and their functions.

As discussed in Section 2.5 of Chapter 2, BullStop was designed to address the flaws identified in existing cyberbullying prevention systems. Specifically, it improves on existing cyberbullying prevention systems in the following ways. Firstly, BullStop is novel in its ability to analyse messages in real-time, protecting users while they use instant messaging applications. Secondly, in existing cyberbullying prevention systems, the classifier used to identify cyberbullying instances is tightly integrated with the rest of system, and it is only through significant effort that the classifier can be modified or a new one introduced; in contrast, BullStop's flexible architecture allows the use of machine learning models in a 'plug and play' manner meaning that different machine learning models can be quickly and easily introduced into the system with minimal or no change. Thirdly, BullStop is novel in its provision of a dedicated and personalised machine learning model for each user. When a user reviews a message and updates the assigned offensive labels, these are marked as potential training data to improve the model. When a sufficient number of training samples have been collected for the user ($n = 1000$) messages, a new instance of the model is created for the user and re-trained using the accumulated training samples. The resulting model is thus a personalised variant of the original classifier that has been specially trained to understand the user's communication patterns and sensibilities. Fourthly, the use of a cloud computing platform (in this case, Microsoft Azure) provides BullStop with the ability to scale to meet the high data traffic demands of popular online social networks. Finally, as reported in Chapters 4 and 5, the app was designed using a collaborative design process with key stakeholders to ensure that the final product represents the needs of the target audience.

6.4 Android Application Framework

The Android mobile operating system is the dominant mobile operating system worldwide, installed on approximately 1.031 billion mobile devices and thus representing 86% of global mobile devices (IDC, 2020). The platform architecture for Android comprises five key layers (as illustrated in FIGURE 6.3). The Application Layer is essentially where all apps are hosted. There are broadly two types of apps, system apps and user apps. System apps are core apps included in the operating system by default that provide core functionalities like SMS messaging, email, internet browsing, contacts and more. User apps are third-party apps installed from an external source like the

Google Play app store. Although included by default, system apps have no special status within the platform; a third-party app can be installed to replace a system app or share responsibility with the system app. In such situations, the third-party app may require elevated privileges to perform the core functions. For example, for BullStop to automatically analyse and delete SMS messages, it required elevated privileges to access SMS messages, effectively becoming the default messaging app. A user app can also instruct system apps to perform core functions on its behalf. For example, the BullStop Help screen allows users to call cyberbullying prevention charities directly from the screen by instructing the Dialer system app to complete the action initiated within BullStop.

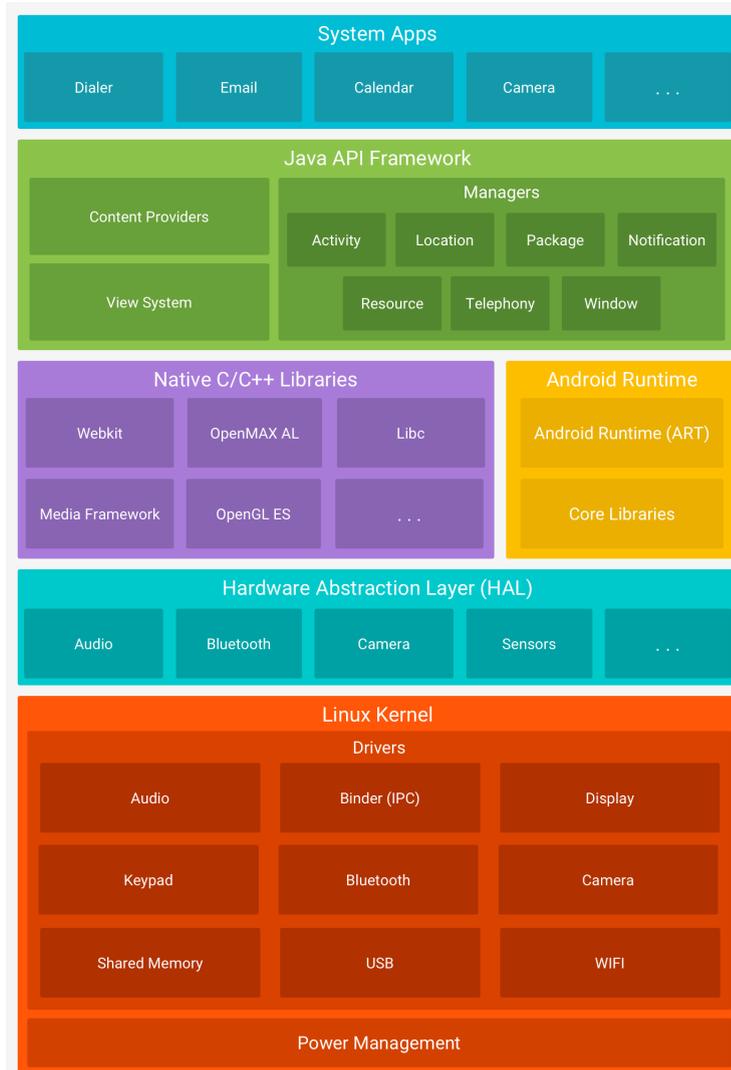


FIGURE 6.3: Android Platform architecture.
Source: developer.android.com.

The Java API Framework provides access to the Java language upon which Android is

built. The Native C/C++ Libraries serve as an interface to the smartphone's hardware drivers. For example, mobile games often use these libraries for drawing and manipulating images on the screen.

The Android Runtime provides the 'virtual container' within which apps run and are allocated mobile phone resources like memory and storage.

The Hardware Abstraction Layer provides access to the smartphone's hardware via an abstracted layer ensuring that an app works in the same way on different devices. For example, an app can access the smartphone's camera in the same way regardless of the type of camera on the phone.

Finally, the Linux Kernel is the foundation of the Android platform and provides low-level functionalities like threading (running multiple apps concurrently) and memory management.

6.5 User Interface

As discussed in Chapter 5, the app's user interface evolved from designs developed using a participatory design approach. The resulting application screens from this work are discussed in the following sections. An overview of the app's screens and their relationship to each other is illustrated in FIGURE 6.4.

6.5.1 End User Licence Agreement

The EULA screen is the first screen displayed to the user after installing the app (see FIGURE 6.5). It presents the app's terms and conditions (see Appendix D.1) to the user, and acceptance of the terms is required to use the app.

6.5.2 Forgot Password

This screen provides a means for users to change their password using the registered email address. The user supplies the required email address and taps the 'RESET PASSWORD' button. An email is then sent to the registered email address with a weblink

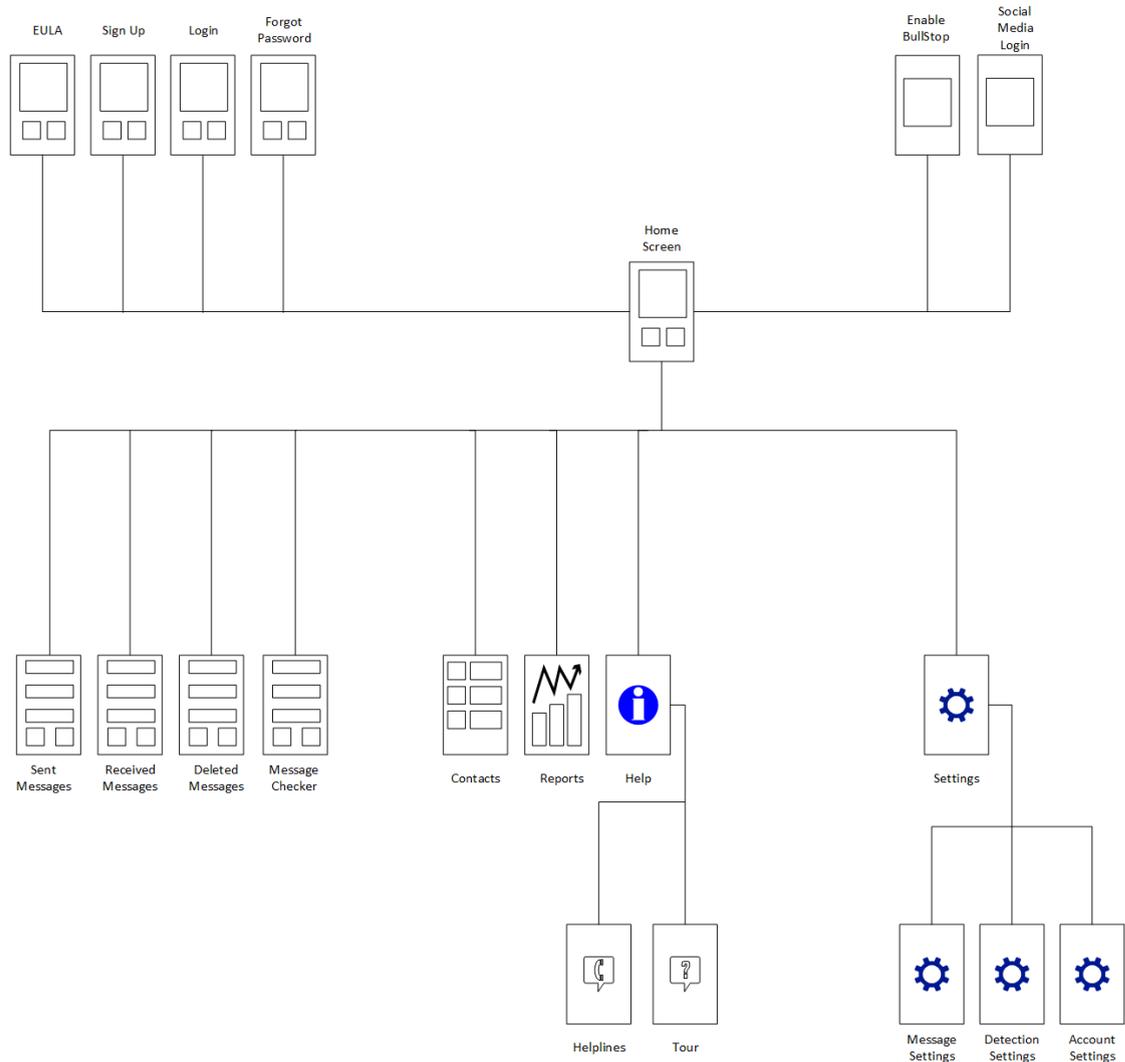


FIGURE 6.4: Navigational map of the app.

to a secure webpage to create a new password. If an unregistered email address is provided, the ‘Change Password’ email will not be sent. The Forgot Password screen is shown in FIGURE 6.6.

6.5.3 Sign In

The Sign In screen is displayed after the user accepts the app’s terms and conditions. Users can login to the app or create a new account via the ‘Sign Up’ link. They can also change their password using the ‘Forgot Password’ link. Rather than storing and managing user accounts locally, the app outsources this functionality to the Okta identity platform.

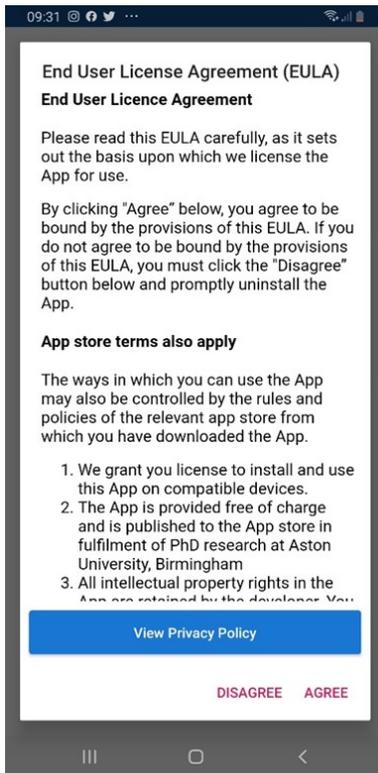


FIGURE 6.5: End User Licence Agreement Screen.

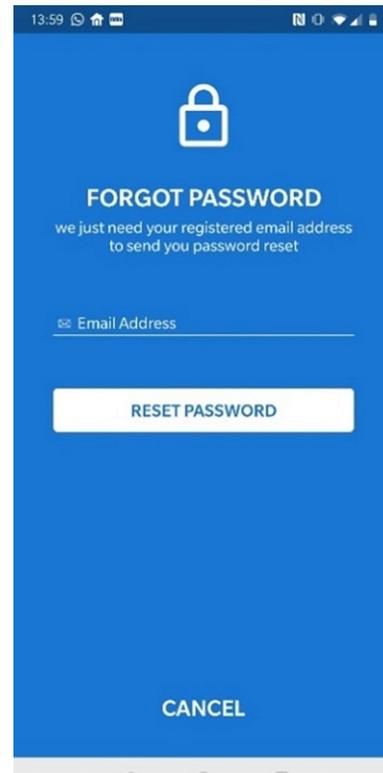


FIGURE 6.6: Forgot Password Screen.

Okta³ is a secure identity and user authentication platform that can be leveraged by mobile and web applications to manage and authenticate users securely. By using Okta, the app can take advantage of standard security features such as password management, email and SMS verification, and malicious threats protection. The app communicates with the Okta platform via APIs. FIGURE 6.7 depicts the Login screen.

6.5.4 Sign Up

A BullStop account is required to use the app; this account is the mechanism used to store personal preferences and is associated with the user's social media accounts. The Sign Up screen allows the user to create a new BullStop account. The Sign Up screen is shown in FIGURE 6.8.

³okta.com

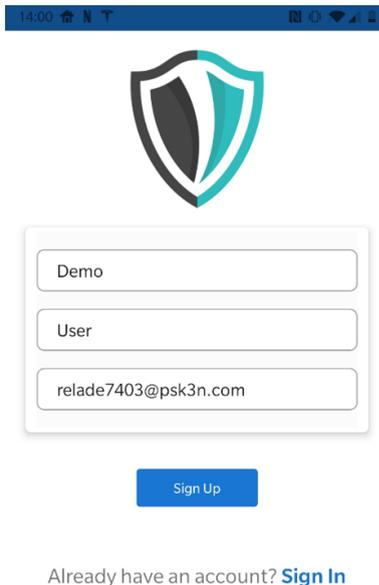


FIGURE 6.7: Sign Up Screen.

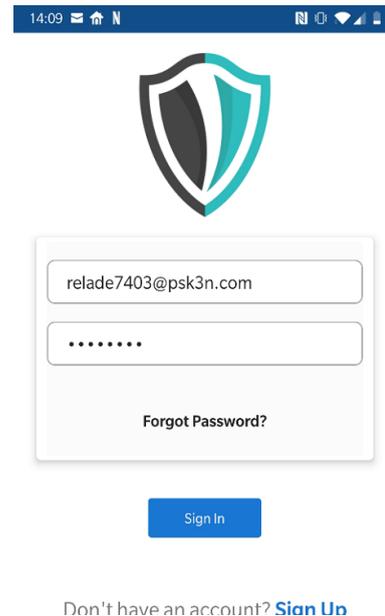


FIGURE 6.8: Sign In Screen.

6.5.5 Home

The Home screen (FIGURE 6.9) provides access to all the app screens available to an authenticated user. An authenticated user is one that has successfully logged in to the app using their BullStop Account.

6.5.6 Social Media Login

The Social Media Login screen allows an authenticated user to link social media accounts to BullStop so that the accounts can be monitored for online abuse. When users tap the Login button, they are securely transferred to the social media website to log in using their social media accounts and provide any required consent. The Social Media Login screen is depicted in FIGURE 6.10.

6.5.7 Enable BullStop

After installation, BullStop defaults to a disabled state. This means that social media and SMS messages are not being monitored. The user is required to enable BullStop to activate message monitoring. If a social media account is not already linked, the user

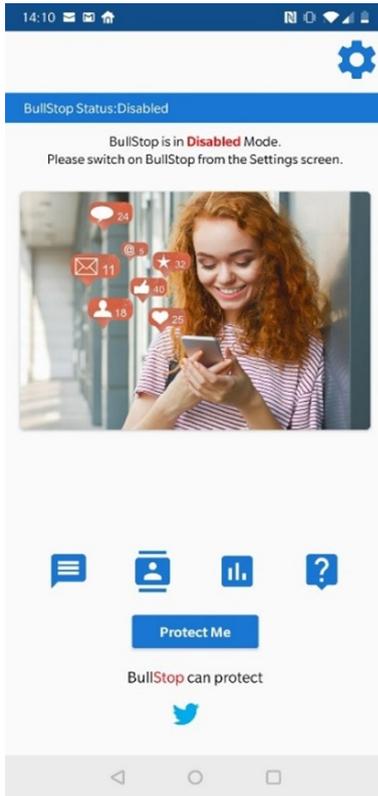


FIGURE 6.9: Home Screen.

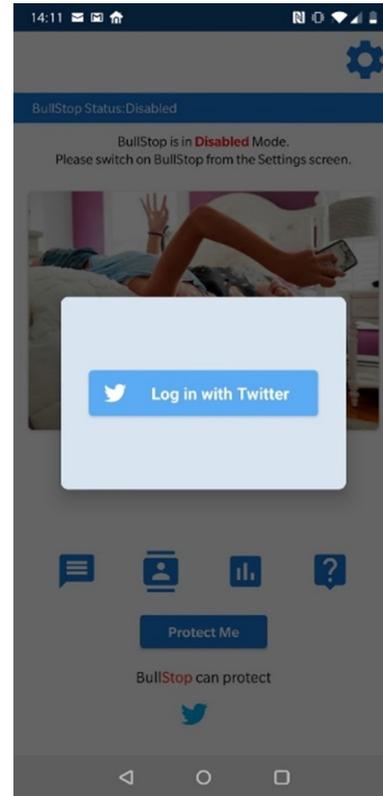


FIGURE 6.10: Login Screen.

is prompted to connect a social media account. The Enable BullStop screen is shown in FIGURE 6.11.

6.5.8 Sent Messages

This screen (FIGURE 6.12) displays messages sent by the user via SMS and social media accounts. All sent messages are automatically analysed for cyberbullying, and offensive content and the offensive labels assigned to the message by the classifier can be viewed by tapping on the message as illustrated in FIGURE 6.13. The assigned labels can be updated by selecting the relevant checkboxes. Such changes will be flagged by the app and saved for use as training data to improve the classifier.

6.5.9 Received Messages

SMS and social media messages received are displayed on the Received Messages screen. Similar to the Sent Messages screen, tapping on the message shows the full

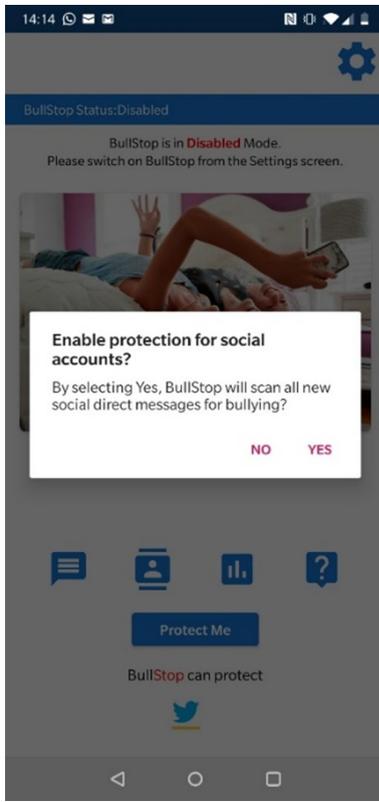


FIGURE 6.11: Enable BullStop Dialog.

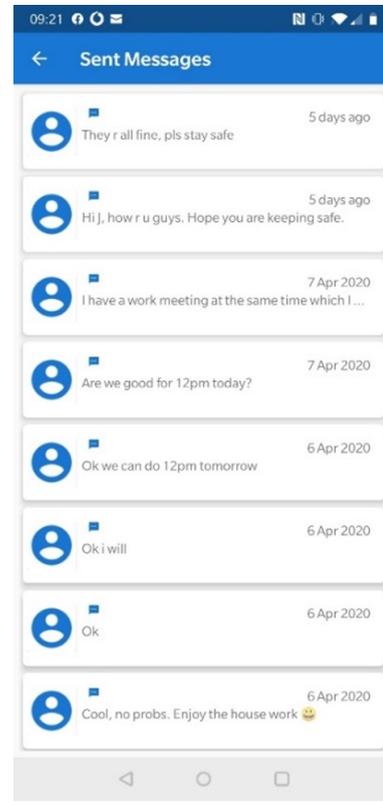


FIGURE 6.12: Sent Messages Screen.

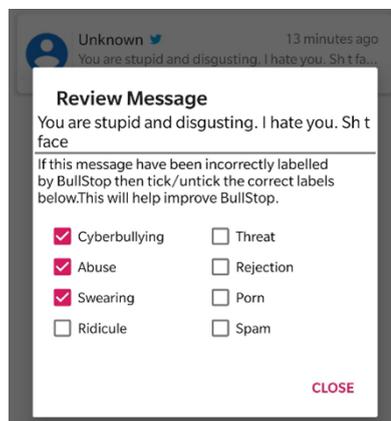


FIGURE 6.13: Message Review Screen.

message content and the offensive content labels assigned to the message by the classifier. As BullStop can automatically delete objectionable messages, the screen only shows messages that have not been deleted by BullStop. The Received Messages screen is shown in FIGURE 6.14.

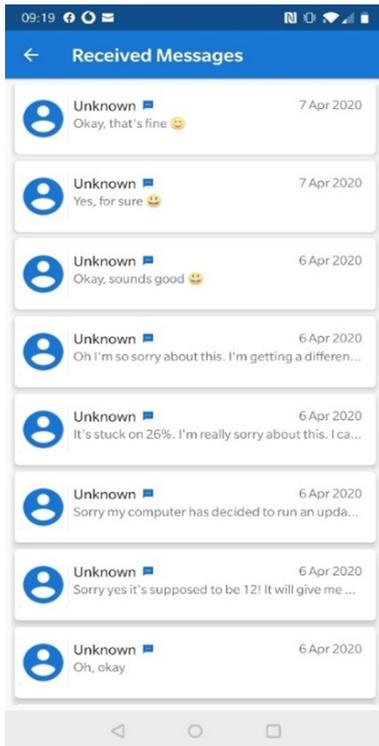


FIGURE 6.14: Received Messages Screen.

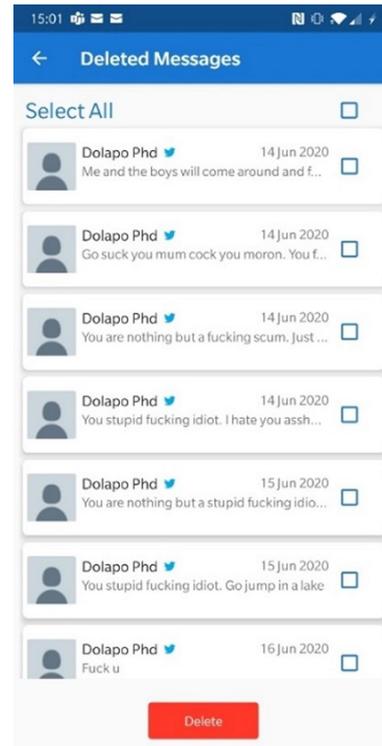


FIGURE 6.15: Deleted Messages Screen.

6.5.10 Deleted Messages

When an offensive message received by the user exceeds the configured offensiveness threshold, it is automatically deleted by BullStop. The message is no longer available on social media or the phone (for SMS messages). BullStop, however, maintains copies of all deleted messages and these are available for viewing via the Deleted Messages screen. Similar to the Sent Messages and Received Messages screens, the labels assigned to a deleted message can be viewed and updated (if desired) by tapping on each message. As the messages have been deleted from the online social network, they cannot be reinstated back to the platform if the user feels they should not have been deleted, but by updating the assigned labels, the user is providing the classifier with additional training data for it to learn more about how to classify such messages in the future. Users can permanently remove

messages using the 'Delete' button or configure BullStop to automatically remove them after a set time. Messages removed from the Deleted Messages screen are permanently removed from BullStop unless the labels have been updated by the user. For messages such as these, they will no longer be displayed in the Deleted Messages screen but will be kept by the system until they have been used to re-train the ML model, they are then permanently removed. FIGURE 6.15 depicts the Deleted Messages screen.

6.5.11 Message Checker

The Message Checker serves as a reflective interface to discourage users from sending inappropriate messages. Users can compose messages and analyse them for offensive content in real-time using the cloud-based classifier. After analysis, relevant offensive labels (if any) are highlighted, and the Send button is disabled to prevent users from sending such messages. If the message is inoffensive, users can send the message to their contacts via SMS or social media. This behaviour of the app (i.e. preventing the sending of offensive messages through the Message Checker) cannot be overridden. The Message Checker screen is shown in FIGURE 6.16.

6.5.12 Manage Contacts

Phone and social media contacts are automatically imported into BullStop and presented via a single view in the Manage Contacts screen (see FIGURE 6.17). This unified view allows users to manage their contacts across multiple social media platforms from within a single screen in BullStop. Contacts can be assigned one of three statuses, namely Blocked, Trusted or Normal. Blocked contacts are prevented from messaging the user while messages from Normal and Trusted contacts are not blocked. The difference between Trusted and Normal contacts is that messages from Trusted contacts are not analysed for offensive content (whereas they are for Normal contacts) as such contacts have been 'whitelisted'. This 'whitelisting' functionality allows a user to receive messages from Trusted contacts regardless of the message content. For example, parents can be marked as Trusted contacts to ensure their messages are never automatically deleted by BullStop.

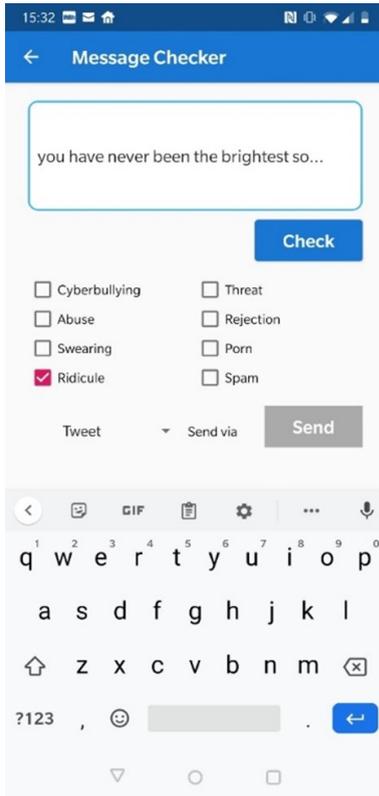


FIGURE 6.16: Message Checker Screen.

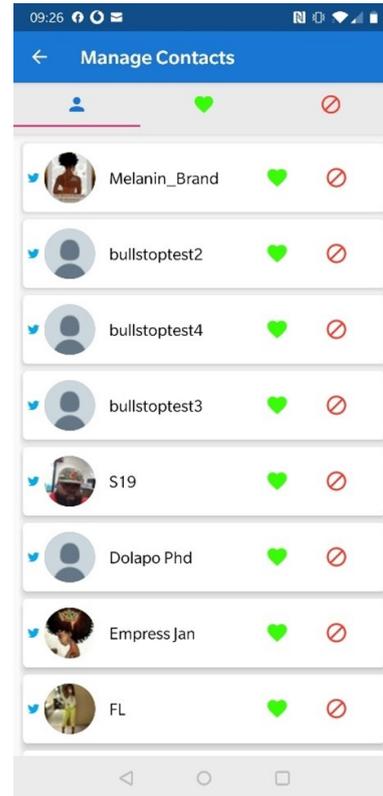


FIGURE 6.17: Manage Contacts Screen.

6.5.13 Report

The Report screen summarises the number of SMS and social media messages deleted and contacts blocked. This is illustrated in FIGURE 6.18.

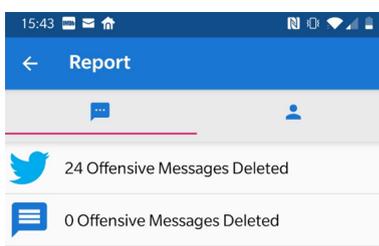


FIGURE 6.18: Report Screen.

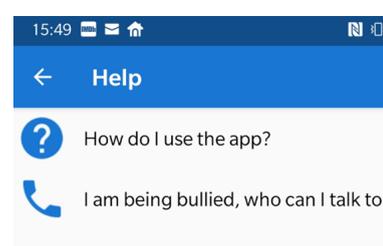


FIGURE 6.19: Help Screen.

6.5.14 Help

The Help screen is a menu screen that provides access to the Helplines and Tour screens, as shown in FIGURE 6.19.

6.5.15 Helplines

The Helplines screen details the telephone numbers for charity organisations that offer assistance to cyberbullying victims. The charities can be called directly from the screen by tapping on the listed telephone number. The Helplines screen is depicted in FIGURE 6.20.



FIGURE 6.20: Helplines Screen.

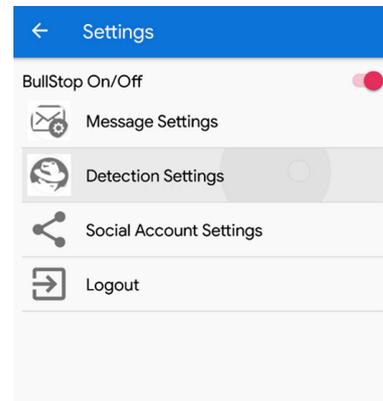


FIGURE 6.21: Settings Screen.

6.5.16 Settings

Similar to the Help screen, the Settings screen (see FIGURE 6.21) serves as a menu providing access to various application settings. The additional screens accessible from the Settings screen are the Message Setting, Detection Setting and Account Setting screens; these are discussed below.

6.5.17 Message Settings

The Message Settings screen allows the user to configure if and how often deleted messages should be permanently removed. The Message Settings screen is displayed in FIGURE 6.22.

6.5.18 Detection Settings

The Detection Settings screen allows the user to control how BullStop manages abusive contacts and sets the offensiveness threshold for messages. As described in Section 6.

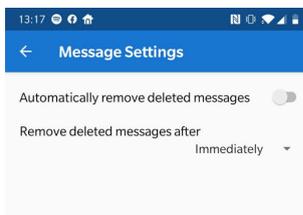


FIGURE 6.22:
Message Settings
Screen.

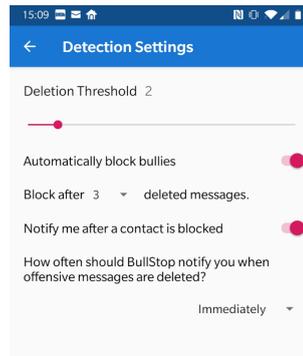


FIGURE 6.23:
Detection Settings
Screen.



FIGURE 6.24:
Social Account
Settings Screen.

2, this determines if the message is deleted by the Marshaller. The Detection Settings screen is depicted in FIGURE 6.23.

6.5.19 Social Account Settings

Users can link BullStop to their social media accounts and authorise the app to monitor the accounts for offensive messages via the Social Account Settings page. The screen is depicted in FIGURE 6.24.

6.5.20 Tour

The Tour screens provide instructions on how to use the mobile app and some of the tour screens are shown in FIGURE 6.25.

6.6 Application Logic

The application logic governs the app's behaviour. It is responsible for screen navigation and manages interaction with the local database and external platforms, including synchronising data with the remote database. It communicates with social media platforms and the cloud backend via the APIs and manages the schedule for data retrieval from online social networks.

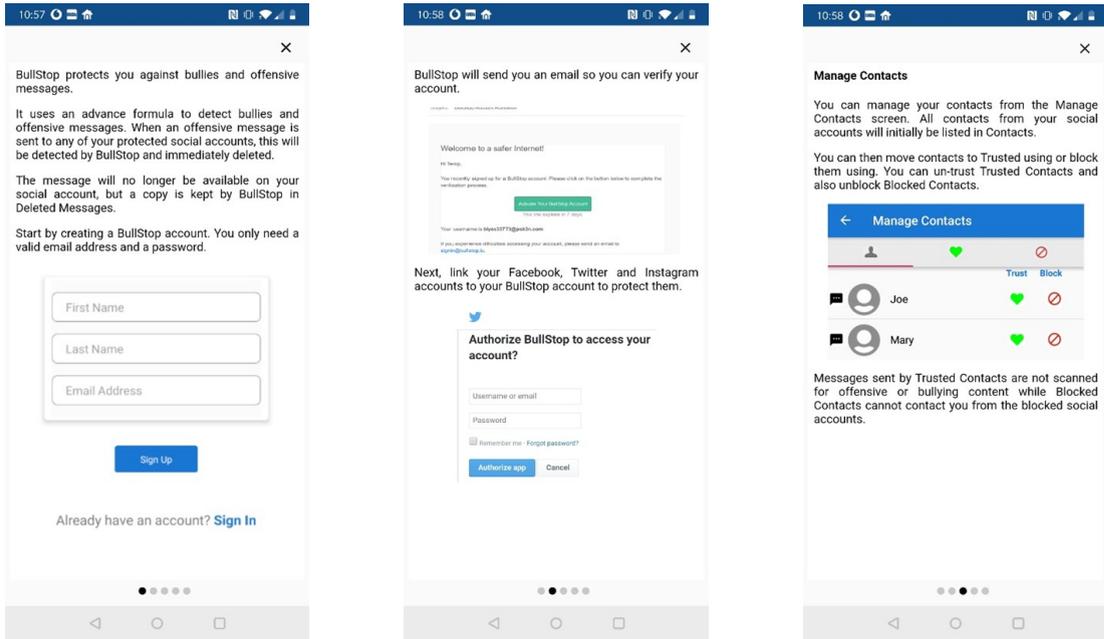


FIGURE 6.25: Tour Screens

6.7 SQLite Database

SQLite is a file-based, open-source Relational Database Management System (RDBMS) designed for portability and reliable performance in low-memory environments. This makes it ideal for use in mobile applications. The database serves as the local data store for messages, contact details, user preferences and other application data. When social media data is retrieved from online social networks, it is first saved in the local database and then synchronised with the cloud database.

6.8 Cloud Backend

The cloud backend is the 'brain' of the system and is responsible for all activities outside of the mobile app. It hosts both the Abuse Detection Module and the Marshaller, two components critical to the entire system's operation. As mentioned in Section 6.2, BullStop was designed to address some of the weaknesses identified in existing cyberbullying prevention systems. Notably, it was designed to be resilient and highly scalable to cope with the high data traffic volume experienced on social media platforms. Furthermore, it 'future-proofs' itself by allowing the use of different machine learning models in a 'plug and play' fashion. These capabilities have been made possible through

a microservices architecture that allows the backend to be implemented as a set of self-contained application services that communicate with each other using a defined messaging protocol.

Offloading the computing-intensive operations (e.g., predicting the offensive labels for messages and re-training the ML models) to the cloud backend made it possible for the mobile app to be implemented as a 'thin' client (i.e., no complex capabilities) that simply serves as a graphical interface for users to interact with the complex functionalities housed in the cloud backend. This significantly reduces the demand on the mobile phone's resources, ensuring that even low-powered mobile devices can run the mobile application. This, however, means that the mobile app cannot operate without being connected to the Internet. As an Internet connection is required to use social media, the app's Internet requirement is not in addition to the basic requirement for online social networking.

Microservices are independently deployable application services that exchange information via messages and are usually supported by a deployment and orchestration framework (Pahl and Jamshidi, 2016). As a software architecture paradigm, microservices have gained tremendous popularity in recent years, emerging as an evolution of the Service-Oriented Architecture (SOA) software movement of the early 2000s. The popularity of cloud computing platforms like Amazon Web Services⁴, Microsoft Azure and Google Cloud Platform⁵, and containerisation technologies such as Kubernetes⁶, Docker⁷ and OpenShift⁸ have contributed to the widespread use of microservices architectures in implementing complex computing systems. Containers provide a means of hosting a microservice within an isolated computing unit, with multiple containers sharing the host computer's resources. When implemented on a cloud computing platform, containers can potentially access an almost limitless source of computing resources along with the ability to rapidly replicate themselves when required. As microservices are designed to be independent entities, the containerisation service (called an orchestrator) can create and maintain multiple instances of the same microservice. If an instance dies, it is simply removed from the cluster and replaced with

⁴aws.amazon.com

⁵cloud.google.com

⁶kubernetes.io

⁷docker.com

⁸openshift.com

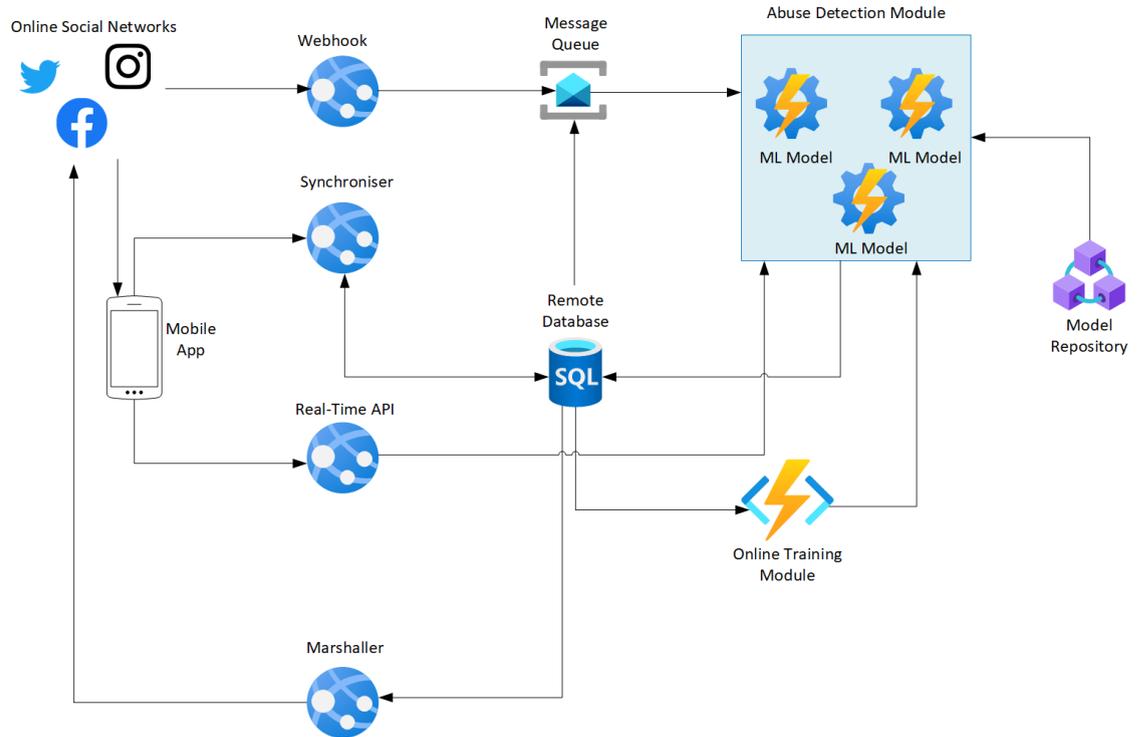


FIGURE 6.26: High Level Architecture of the Cloud Backend

a newly created one. The number of concurrent containers is limited only by the available computing resources.

Microservices are designed to be single-purpose and, as such, provide benefits like the ability to deploy and scale different components of a system independently, resulting in an improvement of the overall reliability and resilience of the system. Social media platforms use microservices to enable them to operate at the scale required to meet the demands of millions of online users and so, by utilising such an architecture, BullStop similarly benefits, providing it with the ability to be deployed at a scale to match that of a modern social media platform. The cloud backend is illustrated in FIGURE 6.26, and its key components are discussed in the following sections.

6.8.1 Webhook

In addition to the mobile app’s regular data retrieval, the Webhook provides an alternative entry point for new data into the system. Unlike the mobile app which retrieves information on a schedule via a pull mechanism, the webhook receives new data via a push from the social networks. When supported by the social media platform, this is an event-triggered

feature that allows the social network to send new data to an external interface when certain events are detected – for example, when a new message is received. The received data is then placed on the Message Queue for onward processing.

6.8.2 Message Queue

While microservices are designed to be independent entities, they do require a means of communicating with one another. The Message Queue provides this means within the cloud backend. It enables the microservices to communicate with each other by exchanging small packets of information in JSON format. When new messages are retrieved from the social network, they are placed on the Message Queue and subsequently picked up for processing by the Abuse Detection Module (ADM). The Message Queue is a persistent and fault-tolerant data store that improves the system's resilience and scalability by ensuring that message analysis does not become a bottleneck for the rest of the system. This is done by monitoring the rate at which messages are enqueued and dequeued. If there are too many messages waiting to be processed, then the processing capacity is increased by creating more ADM instances to handle the increased load. The newly created instances are then destroyed when the message volume drops below a threshold.

6.8.3 Abuse Detection Module (ADM)

The ADM encapsulates the trained machine learning models used to analyse messages for offensive content. It retrieves messages from the Message Queue and predicts the offensive labels for each message. In the current implementation, the ADM includes the following machine learning models: Multinomial Naïve Bayes; Logistic Regression; Linear SVC; BERT; XLNET; RoBERTA; and DISTILBERT (all discussed in Section 2.3.3). RoBERTA, being the best performing model (as discussed in Section 3.4.1), is set as the default classifier, but any other models can equally be used. The model analyses each message and predicts values for the eight offensive labels. The prediction is then outputted in JSON format by the ADM. For example, for a message such as: *You must be gay huh? Why you here? Fag!! And I got 2 TANK YA*. The ADM will return an output similar to the one below:

```
{"Cyberbullying": 1,  
  "Insult": 1,  
  "Profanity": 1,  
  "Sarcasm": 0,  
  "Threat": 0,  
  "Exclusion": 1,  
  "Porn": 0,  
  "Spam": 0}
```

As discussed in Section 6.3, each label is assigned a weight which is used to compute the message's offensiveness score for the message (i.e., the value on which basis a message is deemed offensive and automatically deleted). The weights assigned to each label are as shown in TABLE 6.2. The offensiveness score for the above example message will, therefore, be calculated as 11. Thus, if the user's offensiveness threshold is set to 11 or less, the message will be flagged for deletion. The offensiveness threshold for the system defaults to 2, but this can be changed by the user in the mobile app. To determine the weights, a selection of tweets were shown to the young co-designers to identify the tweets that they would like the app to automatically delete by default (i.e. if the user has not configured the deletion threshold). The offensive labels for these tweets were then predicted using the RoBERTA model. Using the tweets selected for deletion by the participants as a guide, the labels were assigned weights to ensure that all the tweets marked for deletion would breach the threshold and trigger a deletion. As all single-labelled tweets apart from those assigned a single label of 'Porn' or 'Spam' were selected for deletion, the 'Porn' and 'Spam' labels were assigned the lowest weight (i.e., 1) and the default deletion threshold was set as 2 to ensure that tweets that are assigned only the 'Porn' and 'Spam' labels are automatically deleted (as suggested by the participants). 'Sarcasm' and 'Exclusion' being labels often associated with tweets containing indirect forms of online abuse were each assigned weights of 2 and the remaining labels (which were all associated with direct forms of online abuse) were assigned a weight of 3 each.

Label	Weight
Cyberbullying	3
Insult	3
Profanity	3
Sarcasm	2
Threat	3
Exclusion	2
Porn	1
Spam	1

TABLE 6.2: Weights for offensive labels.

6.8.4 Model Repository

The Model Repository is used to store the machine learning models used by the ADM. It is essentially an optimised file store that supports the fast retrieval of the binary model files. When an instance of the ADM is created, the model files are simultaneously loaded into memory. This ensures that the ADM can start processing messages immediately. The Model Repository is critical to supporting the system's ability to use different ML models in a 'plug and play' manner. A *default ML model* configuration variable is used to inform the ADM of the ML model to use for prediction; by simply changing this variable, therefore, a new ML model can be dynamically loaded into memory by the ADM. The ADM checks the value of the *default ML model* variable at regular intervals (currently set to an hour).

6.8.5 Online Training Module

Whenever a user amends the labels assigned to a message, such messages are flagged as training data in the database. When there are a sufficient number of these in the database (referred to as n_{um} and currently set to 1,000), a copy of the model is created and re-trained using the amended labels. This copy of the model then becomes a user's personalised classifier and is exclusively used to analyse the user's messages only. The Online Training Module will continuously re-train this bespoke classifier as more training data becomes available, resulting in a highly personalised model specifically trained for the user using his/her judgement as ground truth.

In order to determine the appropriate value of n_{um} to trigger the initial creation of a personalised classifier, a simulated experiment in which a new instance of the model was

created and re-trained with data including the user-updated labels was conducted with various values of n_{um} . Starting with $n_{um} = 100$, the Online Training Module was configured to create a new personalised model for every hundredth increment of n_{um} and it was not until $n_{um} = 1000$, that a noticeable improvement was detected in the personalised model's predictions compared to the base model. This was then subsequently adopted as the value of n_{um} for the Online Training Module.

6.8.6 Remote Database

The Remote Database is a PostgreSQL database that serves as a remote copy of the mobile app's local SQLite database. PostgreSQL is an open-source RDBMS commonly used for web applications due to its high stability and support for popular programming languages. The Remote Database makes it possible to synchronise a user's data across multiple devices if the same BullStop account is used on multiple mobile devices. Besides holding a synchronised copy of the Local Database, it also stores system-wide configurations such as the default machine learning model and social network APIs' details.

6.8.7 Synchroniser

The Synchroniser performs regular data synchronisation between the local and remote databases. It is implemented as an API that is called regularly by the mobile app after it has retrieved data from online social networks to its local database. New data stored in the local database is replicated to the remote database and vice versa.

6.8.8 Real-Time API

In addition to classifying messages retrieved from the Message Queue, the ADM can perform real-time message analysis. In this mode, messages are received directly through the Real-Time API and are immediately analysed. The mobile app's Message Checker (see Section 6.5.11) uses this API to perform real-time analysis.

6.8.9 Marshaller

The Marshaller coordinates the post-message activities of the platform. It reviews the user's personal preferences to determine what actions to take after a message has been analysed. A user's personal preferences include the offensiveness threshold and post-analysis actions (e.g., deleting offensive messages or blocking abusive users). The Marshaller will then initiate the appropriate action using APIs to communicate with the online social network and the mobile app.

6.9 Limitations and Challenges

Navigating the intricacies of software development for mobile and cloud-based applications while staying true to the original ideals of BullStop presented a number of challenges. These included supporting a diverse range of devices running the Android operating system, online social network API restrictions and the Google Play Store's policies.

As previously noted, as the most popular mobile operating system Android is installed on over a billion mobile devices. Unlike in Apple's iOS mobile ecosystem, whereby devices running iOS versions older than two years are rare, the Android ecosystem includes devices using older versions released many years ago. Furthermore, iOS is only installed on Apple devices with strictly regulated screen sizes, whereas Android runs on a plethora of devices with a wide range of screen sizes. This, therefore, presents a challenge when designing user interfaces for Android mobile apps due to the need to cater to a large number of device screen sizes. To alleviate this issue, Google broadly categorises Android devices into a number of groups based on the screen size (*Small, Normal, Large, Xlarge*) and pixel density (*ldpi, mdpi, tvdpi, hdpi, xhdpi, xxhdpi*), as illustrated in FIGURE 6.27. The size measure relates to the physical dimensions of the mobile phone's screen, while pixel density is the number of pixels per inch of screen, with higher values signifying better display quality. Thus a device designated as *xlarge* and *xxhdpi* represents the largest possible screen size and highest pixel density for an Android device.

To ensure adequate coverage of as many Android device types as possible, BullStop was targeted at devices with screen sizes of *Normal* and *Large*, and pixel densities *hdpi, xhdpi*

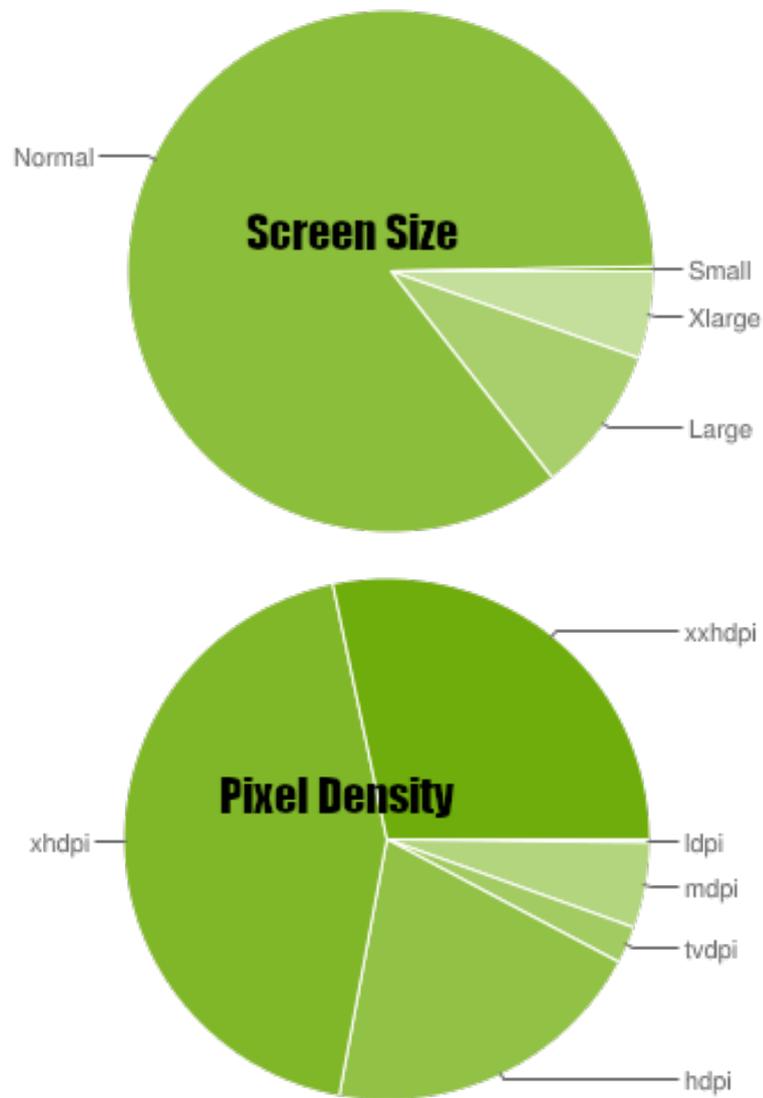


FIGURE 6.27: Android device categorisation based on screen size and pixel density.
Source: developer.android.com

and *xxhdpi*. This approach required significantly more development effort than if a smaller range of size and pixel densities had been targeted, but it ensured that all UI elements are rendered correctly in devices that fall within these groups, with the subset being considered reasonable within the scope of this research agenda. UI elements may be displayed in lower quality for devices outside of these groups, but the app's core functionalities will operate as expected.

The minimum version of the Android operating system required to run BullStop is Android Version 7.0 (released in 2016). This was selected because 75% (approximately 2.25 billion) of all Android devices are using this version or higher. The implication of this

decision, however, is that the app could not take advantage of performance-enhancing features introduced in more recent versions of the Android operating system, forcing the app to run in a 'backward compatible' mode so it can be used on phones running older versions of Android. As devices running older versions of Android are typically older and cheaper, targeting a legacy version of the Android operating system ensures that children with such phones (who may also be from less affluent backgrounds) are not disadvantaged.

As BullStop integrates with online social networks via their APIs, its features are heavily reliant on the functionalities exposed by these APIs. In April 2018, Facebook introduced wide-ranging changes to its APIs and that of Instagram (which is owned by Facebook) as part of its response to its complicity in allowing unfettered access to users' data by Cambridge Analytica (Facebook, 2018). These changes severely restricted access to critical features (from BullStop's perspective) of both the Facebook and Instagram platforms. Specifically, it was no longer possible to retrieve a list of the user's contacts and messages, and operations such as deleting messages and blocking contacts via the APIs were no longer possible. As the original intention was for BullStop to support Facebook, Instagram and Twitter from the onset, these API restrictions severely hindered the app's ability to integrate with Facebook and Instagram. The decision was made, therefore, to remove support for both Facebook and Instagram leaving Twitter as the only social media platform currently supported by the app for this research. While the proportion of Twitter users aged 13 – 17 years and 18 – 24 years are about 9.1% and 21.6% respectively (Statista, 2020b), it is not believed that having Twitter as the only social media platform that can be used with the app significantly limits its ability to engage the target audience for research purposes as the population of these age groups on Facebook (13 – 17 years: 5.8%, 18 – 24 years: 23.5%) (Statista, 2020a) and Instagram (13 – 17 years: 7.1%, 18 – 24 years: 29.6%) (Statista, 2020c) are similar. It is hoped that the reach of BullStop can be expanded in the future but, within the context of a research agenda, the use of Twitter would still permit credible exploration and validation of the research concepts.

The Google Play Store is the official store for Android applications and is the main distribution channel for Android apps. Google recently changed the policies regarding sensitive permissions required by apps (Google, 2020), and one of these changes relates to necessary permissions to view and manage SMS messages. The change required an

app to become the phone's default messaging application to view and compose SMS messages. This effectively means that for BullStop to access the SMS messages on the phone, it has to become the default messaging application. As BullStop is not designed to be a messaging application, and efforts to agree on a workaround with Google were unsuccessful, the ability to manage SMS messages and phone contacts was temporarily disabled in BullStop to allow a successful submission to the Google Play Store and the research to continue. This feature will be enabled in a future release of the app, and discussions are ongoing with Google to facilitate this.

The above challenges highlighted some of the practical issues affecting the development of impactful cyberbullying prevention tools. Interestingly, the majority of these challenges are not technological; rather, they are due to the policies and actions of technology and social media corporations. Facebook's near-total restriction of its APIs is a gross overreaction on the organisation's part in response to data misappropriation practices that they facilitated – a decision that has rendered many legitimate uses of these APIs impossible. There are several ways in which controlled and managed access to the APIs could have been implemented instead of a blanket removal of crucial features. Undoubtedly, personal data is precious and should be protected at all times; however, many mechanisms are available (some of which were developed by Facebook) that allow users to provide consent and authorise applications that they use to access their data securely. While it is disappointing that Facebook's policy change negatively impacted the reach of BullStop, it is hoped that this can be addressed in the future and it is not felt it impeded the research focus being reported here.

Google and Apple have been criticised in the past for their overbearing app stores' policies, with accusations of innovation stifling levelled at both corporations (Ranger, 2016; Low, 2018; Solsman, 2019; Warren, 2020; Aten, 2020). The experience gained in publishing BullStop on the Google Play Store would certainly give credence to this. For example, despite several attempts (spanning four months) to convince Google of BullStop's requirement for requesting SMS messages access, the only way to get the app approved was by removing this feature, thus limiting one of the app's core functions.

While these issues may limit some of BullStop's features, the app still fulfils its core objective of providing a means for potential cyberbullying victims to protect themselves

from online abuse. Future enhancements to the app will target reducing the impact of some of these limitations, and these are discussed in more detail in Chapter 9.

6.9.1 Summary

BullStop has been developed with the principal objective of providing cyberbullying victims with a means of protecting themselves from online abuse. As a cyberbullying detection and prevention application, it has been designed to overcome identified shortcomings of existing tools. It is composed of two sub-systems – an Android mobile application and a cloud backend. The mobile application is the key interface for users to interact with the system and features a novel UI designed collaboratively with stakeholders using a participatory design methodology. The cloud backend utilises industry-leading technologies such as containerisation, cloud computing and microservices to improve the system's resilience and scalability, thus ensuring that it can cope with the high data traffic volume of modern online social networks.

Unlike existing cyberbullying prevention systems that are based on a specific classifier and are unable to change machine learning models without significant effort, BullStop is designed to be highly flexible and can use different machine learning models in a 'plug and play' manner. This novel functionality improves the system's versatility and has never been used before in a cyberbullying prevention tool. Furthermore, by hosting the classifier in the cloud backend, the system is able to offload the intensive computational demands of running a machine learning model to the cloud platform, ensuring that the typically minimal computing resources available on mobile devices do not negatively affect the performance of the app. Unlike existing work on cyberbullying prevention tools, BullStop is not merely an experimental prototype; it is a viable cyberbullying prevention tool that is ready and available for use and one that has been openly evaluated on the Google Play Store by hundreds of users.

Some practical challenges encountered in the app's development phases enforced some compromises in the released version of the app. While these compromises do not detract the app from its key objectives, they nevertheless reduce the available app's features. Many of these compromises were, however, due to the operating practices of technology and social media companies rather than constraints of technology. This highlights an

urgent need for these corporations to remove practices that impede innovative use of their platforms, especially those intended for research and altruistic purposes.

Chapter 7: Evaluation of the BullStop Mobile Application

7.1 Introduction

Utilising user-centred and participatory design approaches helped identify the pertinent issues for stakeholders regarding how cyberbullying is being perpetrated, their opinions on current prevention strategies, and how the innovative mobile application (Bullstop) should be implemented to provide the maximum benefits for users. This chapter details the evaluation studies conducted to explore the responsiveness, scalability and perceived usability of the BullStop app. Sections 7.2 and 7.3 reports on the computer-based experiments performed to evaluate the system's responsiveness and scalability. The ways by which the system mitigates obsolescence by design is considered in section 7.4. To access the mobile app's acceptability amongst the target audience, two types of evaluation study were conducted, namely 'lab'-based and field-based evaluations. Section 7.5 reports on the 'lab'-based evaluation study, which comprised a series of exploratory sessions conducted with parents, children, clinicians, and law enforcement during which they were provided with the mobile app and their initial impressions solicited. In the field-based evaluation study, the app was made available to the public via the Google Play Store and beta users were essentially invited to complete an online questionnaire about their experiences with the app: this is discussed in Section 7.6. The chapter concludes with a summary in Section 7.7.

7.2 Evaluation of the System's Responsiveness

The Real-Time API was the mechanism chosen to evaluate the system's responsiveness. The Real-Time API is an API interface to the ADM (Abuse Detection Module) that accepts text input, predicts the offensive labels for the text and returns the output immediately in JSON format (see Section 6.8.8 in Chapter 6). It is, therefore, an appropriate component of the system to demonstrate this capability of the application. Postman¹, an API client tool used to send web requests to an API, was used to perform the evaluation experiments. Multiple messages to classify were provided in a text file, and Postman was configured to iterate through the file and send the messages to the API. The messages comprised a random collection of tweets retrieved via the Tweeting streaming API using the procedure described in Section 3.3.4. The time taken by the API to return the prediction is the 'response time', and this was recorded for each message. Three experiments sending varying numbers of tweets (100, 500, 1000) to the API were conducted, and the average response time for each experiment was calculated. For each experiment, the tweets were provided in a separate text file to Postman for processing. While there is no officially agreed value for an adequate response time for an API, and acceptable values vary depending on the operations being performed by the API, a value of 1 – 3 seconds is widely regarded as an adequate response time for an API (Nielsen, 1994; StackOverflow, 2008); this was therefore adopted as the expected range of values to demonstrate responsiveness. The same system configuration used in the field-based evaluation study (see Section 7.5) was used. In this configuration, the system can access 14GB of memory and four 2.4GHz virtual Intel processors. The results of the experiments are presented in TABLE 7.1.

Experiment	1 (100 tweets)	2 (500 tweets)	3 (1000 tweets)
Average Response Time (sec.)	1.068 ± 0.075	1.013 ± 0.049	1.039 ± 0.038
[95% Confidence Interval]			

TABLE 7.1: Results of Responsiveness Evaluation Experiments.

As shown in the table above, an average response time of about 1 second was calculated in all three experiments, placing the average response time within the expected range and thus demonstrating the system's responsiveness. It could perhaps be argued that instead of the computed average response time, users' perception of the system's responsiveness

¹postman.com

is a better indicator for evaluating this attribute; as such, this is considered in Section 7.5.6.4 as part of the field-based evaluation study.

7.3 Evaluation of the System's Scalability

As mentioned in Chapter 1, scalability relates to a system's ability to increase its capacity by increasing the amount of computing resources in use. To evaluate the system's scalability, the ADM was utilised in its asynchronous operation mode (see Section 6.8.3). The BullStop system manages the amount of computing resources it requires by monitoring the rate at which messages are enqueued and dequeued. If the enqueueing rate exceeds that of dequeuing, new ADM instances are created to handle the additional load. This rate is referred to as Message Consumption Rate (e_dC) and denoted by the equation:

$${}^e_dC = \frac{M_e}{M_d}$$

which is simply the division of the number of messages enqueued per second M_e by the number of messages dequeued per second M_d . e_dC can therefore be used as a measure of the system's responsiveness as ${}^e_dC > 1$ implies that the system is under heavy load and additional instances of the ADM should be created to reduce the value of e_dC closer to 1. At ${}^e_dC = 1$, the system is at a perfect equilibrium with the exact amount of resources required to process messages provisioned while ${}^e_dC < 1$ means that the system has over-provisioned resources.

An experiment was performed to assess the scalability of Bullstop by artificially enqueueing messages onto the Message Queue using a computer program developed to test the system by putting messages onto the queue at an approximate rate of 20 messages/sec for three hours and monitoring the values of (e_dC) as the system's load was increased. FIGURE 7.1 illustrates the results of the evaluation experiment and shows (e_dC) increasing as M_e exceeds M_d during the initial period when the system struggled to cope with the increasing number of messages. New instances of the ADM were therefore created to handle the additional load, bringing the value of (e_dC) closer to 1 and thus demonstrating the system's ability to scale and acquire additional computing resource as required. This is a significant differentiator between the BullStop application and existing cyberbullying

prevention systems as it provides BullStop with a theoretical ability to match the scale of modern social networking platforms.

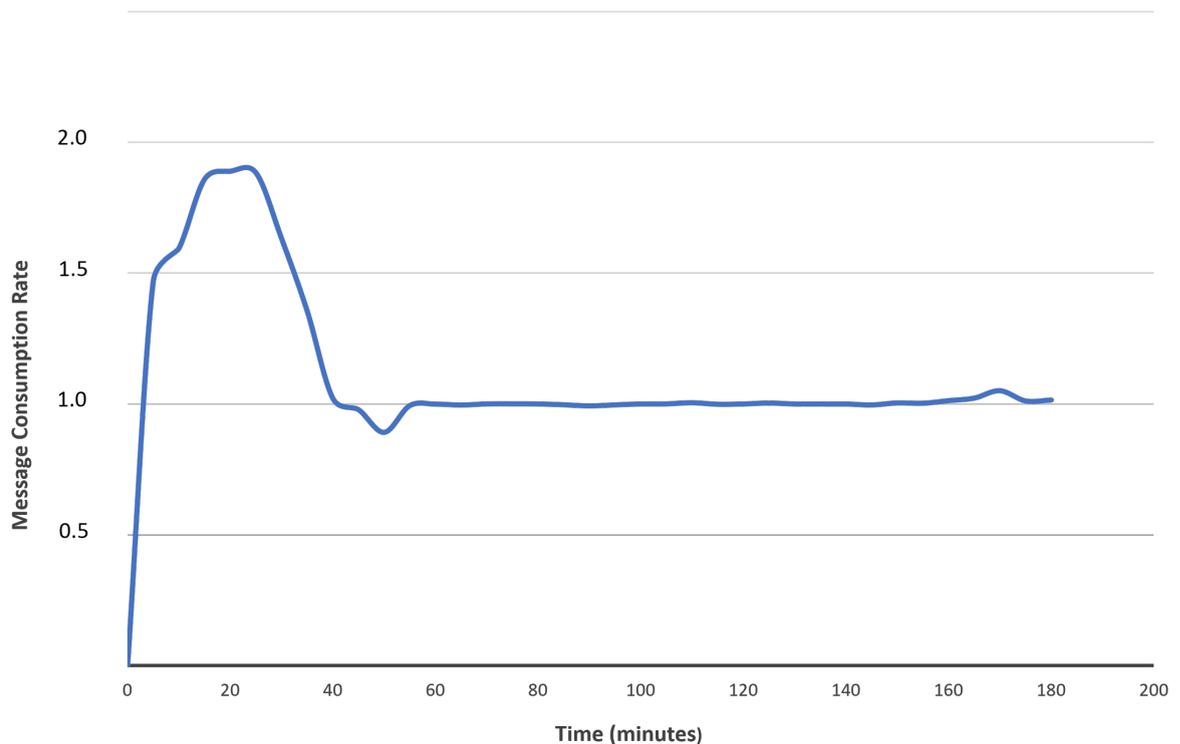


FIGURE 7.1: Graph illustrating Message Consumption Rate under artificially-induced load.

7.4 Mitigating System's Obsolescence

Obsolescence mitigation is built into the design of BullStop. This is achieved via two key means; allowing the use of different ML models in a dynamic manner and the system's ability to generate a personalised classifier for each user. As previously discussed in Chapter 6, the choice of the ML model used for predictions is governed by a configurable value which when updated allows the system to retrieve a different model to use as the classifier from the model repository. This design mitigates system's obsolescence as it decouples the classifier from the rest of the system thereby allowing the use of newer ML models as they become available. In addition, by generating personalised classifiers, the system provides a means to continually learn from the users ensuring that the ML model evolves as the user interacts with the system.

Finally, transfer learning - the use of a source domain and learning task to improve a predictive function for a different target domain (Pan and Yang, 2009), has the potential to be used as an additional means for mitigating obsolescence as knowledge from a new domain can be introduced into the system in this way thus providing a means for the system to expand beyond abuse detection and prevention to areas such as preventing sexual grooming and stalking. In incorporating obsolescence mitigation into the design, BullStop is a very unique tool as designing to address obsolescence in an online abuse detection or prevention tool is a novel endeavour that has not been previously attempted.

7.5 ‘Lab’-Based Evaluation of the Application’s Acceptability to Users

This section reports on the activities conducted during the ‘lab’-based evaluation study.

7.5.1 Study Design

The ‘lab’-based evaluation study was aimed at gaining insight into first-time users’ impressions of the app and garnering detailed feedback about how well the app meets the stakeholders’ expectations. It was essentially a controlled study using lab-based techniques but conducted in more comfortable settings for the benefit of users. Building on the rapport previously established with the focus group participants, the researcher was able to arrange the evaluation sessions quickly and with relative ease. In addition, the participants were genuinely interested in exploring and experiencing the mobile app, which had been based on their input. After confirming their interest in participation by replying to the invitation email, a telephone conversation took place between the researcher and each adult participant. This call was to provide an overview of the evaluation session and to answer any questions that the participants had. Seven sessions took place at the home of participants, and two were held over coffee at a café. All participants (including the young participants) were entitled to a £10 Amazon voucher, but all the adult participants waived this and received no compensation for their time.

Participants were provided with an Android smartphone with the BullStop app pre-installed but not configured. To remove the need for participants to share their personal data and

online social networks information for evaluation purposes, dummy Twitter and BullStop accounts were created for the use of participants. Information on how to configure the app and a list of tasks they were asked to perform were provided to participants (see Appendix E.1). Each session lasted about two hours, the first hour of which involved participants familiarising themselves with the app and performing the tasks detailed in the instruction sheets. They were also encouraged to explore the app and discuss their observations. Afterwards, the researcher interviewed the participants about their observations from their use of the app (a copy of the interview guide is presented in Appendix E.2).

7.5.2 Participant Recruitment

Ethics approval for the 'lab'-based evaluation study was granted by the University Ethics Committee (see Appendix E.3). Participants for the 'lab'-based evaluation study were recruited from the previous focus groups (see Section 4.2.1 in Chapter 4) as well as the original pool from which the focus group participants were sourced. All participants were invited via email (see Appendix E.4 and E.5) with attached Participant Information Sheets (see Appendix E.6, E.7, E.8) and Consent Forms (see Appendix E.9, E.10, E.11). The study consisted of nine evaluation sessions between January and February, 2020. Five sessions were designed as paired exploration sessions with a parent and child, and four were individual evaluation sessions of the mobile app with the psychiatrist, child psychologist, general practitioner (GP) and law enforcement officer who took part in the focus group study (discussed in Section 4.2). An overview of the study's sessions and participants is presented in TABLE 7.2.

7.5.3 Findings and Discussion

Thematic analysis using the procedure established in Section 4.2.2 was performed on the transcripts, and six key themes were uncovered. These, along with the resulting codes, are presented in Appendix E.12 and discussed in the following subsections.

Session	Venue	ID	Role	Gender	Age (Child only)	Original Focus Group Participant?
1	Participant's Home	P1	Parent	Female	N/A	No
		P2	Child	Female	11	No
2	Participant's Home	P3	Parent	Male	N/A	Yes
		P4	Child	Female	16	No
3	Participant's Home	P5	Parent	Male	N/A	Yes
		P6	Child	Male	13	No
4	Participant's Home	P7	Parent	Female	N/A	Yes
		P8	Child	Male	15	No
5	Participant's Home	P9	Parent	Female	N/A	No
		P10	Child	Male	17	No
6	Participant's Home	P11	Psychiatrist	Female	N/A	Yes
7	Cafe	P12	Law Enforcement Officer	Male	N/A	Yes
8	Participant's Home	P13	General Practitioner	Female	N/A	Yes
9	Cafe	P14	Child Psychologist	Female	N/A	Yes

TABLE 7.2: Overview of evaluation study participants

7.5.3.1 Easy to Use

Participants' initial impressions of the app were that the UI is simple and easy to use. They were very positive about their experience setting up the app for the first time. The law enforcement officer noted that:

"It's easy to set up an account initially, and then to log in, it's quite easy and quick. There's not a lot of confusing stuff, and it's really quick to create an account. That's very good". [Law Enforcement Officer]

The Child Psychologist commented that the app was "easy to set up" and that "it connects well with Twitter". It was particularly reassuring that one of the parents that confessed to not being "good with technology" reported a similarly positive experience with the app:

"I think it was quite good for a techy challenged person, easy to use, and straight forward, and not too busy". [Parent]

The perceived learnability of a system is often a key determinant of users' satisfaction (Calisir and Calisir, 2004). The provision of help and documentation to assist with the

system's use is, therefore, a crucial element in increasing an application's learnability and appeal to users. The app's tour is the embodiment of this principle, and with remarks such as those below, the tour has certainly been beneficial in improving the overall appeal of the mobile app.

"I like the tour. I generally struggle with apps or anything techy, but with a few screens of the tour, I understood the app completely". [Psychiatrist]

"My initial impression was very good. You start the app; you can see the tour. As a user, or even as perhaps a guardian, I can go through the tour and understand very quickly what the app is trying to achieve. That is a very good start because some apps can be confusing when you first open them, and that makes this good". [Law Enforcement Officer]

7.5.3.2 Well Designed

Participants commended the app's user interface with many saying that it is well designed and of equal or better quality to apps that they have purchased from the app store, as illustrated by the quotes below:

"It certainly looks like a paid app I would have downloaded from the app store". [Young Participant, 17]

"It looks very professional. Like apps [that] I have paid for before". [GP]

An application's user interface plays a significant role in users' satisfaction (Hiltz and Johnson, 1990) and, for BullStop, the need to satisfy both primary (young people) and secondary (adult stakeholders) audiences introduces additional complexities. In addition to appealing to young people and their parents to gain their acceptance, the app must equally demonstrate its utility to the professional stakeholders so they can promote the app amongst the primary audience via recommendations and word of mouth. On the strength of the comments below from the young participants, the app's interface meets their expectations:

"I like the design and the colours. It's very subtle, not in your face". [Young Participant, 13]

"I like the pictures of the kids on the home page. And the settings, it's straightforward." [Young Participant, 11]

"It's simple. It's like Twitter but also different. I get it from the beginning. I like the settings". [Young Participant, 15]

The parents and clinicians views were also aligned with those of the young participants, as evidenced by the statements below:

"I like the simplicity of the UI. I 'don't like cluttered apps with too many things going on the screen". [Parent]

"I like the way everything is set out". [Child Psychologist]

7.5.3.3 Appropriate Branding

A key finding from the PD sessions was the significant role played by an app's brand in attracting potential users to the app on the app store. Often, an app's name and logo are the only means of differentiating it from millions of other apps on the app store, and both must capture the interest of potential end-users from the onset. An app's logo personifies its image and participants' reactions to it provided us with a measure of its suitability to represent the app's brand. Participants liked Bullstop's logo; it was immensely popular amongst the young participants and clinicians who appreciated the symbolism represented by the shield image used in the logo as exemplified by their comments below:

"I like the shield. It has a strong message about what it is stopping". [GP]

"I like the logo too. It's a strong image. Secure and safe. It's good". [Law Enforcement Officer]

"I like the shield because 'it's like you have a knight protecting you". [Young Participant, 11]

"Actually, that logo is very apt. I like it." [Psychiatrist]

The app's name was similarly well-liked. The child psychologist remarked:

"I like the name; BullStop. When I heard the name, it resonates because cyberbullying and stopping cyberbullying. I just like the name. In short, I think its children friendly". [Child Psychologist]

One of the parents also said:

"The name is very clever. I got it immediately". [Parent]

The young participants said the name was "cool" and "clever". In clarifying these comments, one of the young participants said:

“You don’t want an app that says ‘Cyberbully Stopper’ or something like that, that’s just in your face. BullStop is like Reddit or Twitter. You know what it’s all about from the name, but they don’t make a fuss about it. It’s just there. BullStop. That’s it.” [Young Participant, 16]

This is very encouraging for the app’s prospects amongst the target audience. Additionally, participants remarked that the app projects an image of a “*safe and secure environment*” in their minds, which is a useful quality for an app that requires users to entrust their personal and social media data to it.

7.5.3.4 Good Overall Performance

Alongside the key dimensions of the user interface design, learnability and relevance, a system’s performance is another determinant of end-user satisfaction (Hiltz and Johnson, 1990; Gatian, 1994). It provides the user with a measure of the application’s suitability for its intended purpose. Participants were impressed with how responsive the app’s interface was to touch commands and the speed of real-time predictions by the Message Checker. One participant said of the Message Checker:

“It’s quite fast. I timed the message checker a few times, and it was like 1 or 2 seconds. That’s impressive”. [Young Participant, 17]

Another passed the following comment about navigating between the app’s screens:

“Going from screen to screen, it was quite smooth”. [Law Enforcement Officer]

With regards to the app’s offensive content detection, participants were generally satisfied with the app’s ability to detect and delete offensive messages. Some, however, reported that the app sometimes generalises on the type of abusive content detected, ignoring the finer details of subtle or more complex messages. For example, one participant said that the app is “*quite good at picking abuse and bullying, but it didn’t get the threatening messages*”. This was found to be due to the low number of threatening tweets within the dataset (as identified in Chapter 3) used to train the app’s ML model negatively impacting its ability to detect these types of messages.

7.5.3.5 User Empowerment

Participants were impressed with the level of control afforded users by the mobile app. They said that this contributes to the empowerment of cyberbullying victims by providing

them with a means to control how people interact with them online. The young participants particularly liked the ability to manage contacts from different social media platforms from within the app. Some of them said:

"I like that you can make contacts trusted or blocked. And that it can automatically block people if they are being offensive". [Young Participant, 15]

"It's quite handy to manage friends from different social networks in one place". [Young Participant, 17]

They also liked that they could control the app's sensitivity to offensive messages as shown by the statements below:

"It's good that with the app, I can control what kind of messages I receive". [Young Participant, 11]

"The fact that you can adjust the app settings is very good. That way you can control what type of messages are deleted". [Young Participant, 16]

Bullying (in all its forms) is an assertion of interpersonal power by the bully over the victim (Craig and Harel, 2001) and the clinicians were of the opinion that, by allowing users significant control over its configurations, the app is disrupting this relationship thus allowing victims to wrestle back control from potential online abusers. The psychiatrist said:

"A thing like this, it's putting the young person in control, and giving them some trust, and saying, 'Actually, you know what? You've got this, you can take care of yourself'". [Psychiatrist]

Likewise, the child psychologist shared a similar opinion and explained that, in providing young people with the means to filter abusive messages rather than forcing them to avoid the Internet altogether, the app is an improvement on the advice typically offered to parents on mitigating cyberbullying:

"What we say to parents is instead of having this effect on your child, why 'don't you switch off the Internet at home, take off the device from the child. But in a way that can sometimes contribute to the feeling of powerless in young people because now they have lost their phones and Internet because of this, so it is key to demonstrate to them that they do have the power to fight this and I think this app can provide that". [Child Psychologist]

A similar view was echoed by the GP, who said:

"I think having something like this gives the child some measure of control back, and I think that's key in fighting bullying". [GP]

As is to be expected, the young people welcomed the absence of parental monitoring features in the app. Interestingly, and in line with the focus groups' findings, the adult participants were also not in favour of including parental monitoring facilities in the app. They believe that providing young people with a tool to manage their online interactions is a superior tactic than monitoring them. Some of their comments in support of this view include:

"I 'don't think any child will install an app if they know it allows their parents access to what they are doing on social media. So, I think the decision not to include parental monitoring in the app is the right one". [Law Enforcement Officer]

"I think it's good to give children some control instead of sending parents copies of their messages. No child likes that, and they probably won't use the app". [Parent]

"I think not having a parents', or companion app is very good and should be used as a selling point of this app, and "that's useful because you want young people to use it, otherwise, there's no point for it in the first place. Teenagers will never use it if they know 'there's parent supervision in the app". [Law Enforcement Officer]

7.5.3.6 Reflective and Educative

The Message Checker emerged as one of the favourite aspects of the app for many participants. The younger participants welcomed the idea of being able to check their messages for appropriateness before sending. The youngest participant (aged 11 years) in the study said this when questioned about her favourite feature of the app:

"I think the one when I was sending a text, and then it checked for me because as much as I 'don't want to be hurt, I 'don't want to hurt people either. Also, if people don't want to get in trouble for sending something, they can use the app to tell them if what they are sending is bad". [Young Participant, 11]

Other participants expressed similar sentiments about the Message Checker. One parent said:

“The best thing for me about the app is I think it tells you to stop and think before sending a message, which I think young people struggle with. Some people might not have the intention to bully, but at the end of the day, what you have said has gone a long way to hurt or to bully another child. The fact that it has a stop and think section, which for any child who does not directly want to bully or does not have the intention of bullying, that is positive”. [Parent]

This statement perfectly describes the reflective ideals of BullStop. The Message Checker’s purpose is to educate and allow users (particularly the younger ones) to reflect on the potential impact of how they communicate and to serve as a “*second pair of eyes*” when composing messages. Overall, it can be seen that participants felt that the app is not only a useful tool for cyberbullying victims but that it also serves as a learning tool, educating young people on how to communicate appropriately so as not to offend others.

7.5.3.7 Useful and Unique

In concluding the evaluation sessions, participants were asked about their overall impressions of the app and their intention to use or recommend the app to others. While none of the young participants involved in the ‘lab’-based evaluation study reported suffering online abuse, they said if they were, they would use the app as typified by the comments below:

“Overall, I think that it’s a very good app. I would definitely use it”. [Young Participant, 13]

“I’m not being cyberbullied, but if I am, I will definitely use the app”. [Young Participant, 15]

“Yes, I guess if I have been harassed. Maybe not so much now, but when I was younger, I can see myself installing something like this if it was available then”. [Young Participant, 17]

Some of them also said they would recommend the app to friends:

“I really like it, and I know a few people that I think this will be good for.” [Young Participant, 17]

“It’s pretty good, the BullStop app. It protects young people from the dangers of social media. I will tell my friends about it”. [Young Participant, 11]

One young participant said she “*would definitely use it*”. The parents were equally complimentary about BullStop:

“For my kids, I would suggest to them that why ‘don’t you just filter all your messages through this app because you might be chatting with a friend and jokingly say something offensive or vice versa but an app like’s this just act as a filter, a shield”. [Parent]

“This is an app I can tell my children about. To be honest, I know adults that can benefit from an app like this. What you are doing is life-changing really”. [Parent]

The clinicians and law enforcement officer were very optimistic about the app and its prospects. They said it is unique and would gladly recommend the app to young people who can benefit from it. In their words:

“Definitely, I can recommend this [app] to them because it is putting the advice we provide into effect; block offensive contacts, review connection requests, etc., these are all the practical things we tell the parents to do”. [GP]

“It is within our professional capacity to recommend apps like this because we know that these are tools that can help in their everyday lives”. [Child Psychologist]

“If someone is being bullied online, I would definitely recommend this app to them, and if this tool is publicly available, it’s something the police can suggest to people as a safeguarding tool”. [Law Enforcement Officer]

“I’m not aware in my professional capacity of any software or app that particularly addresses cyberbullying. This is something we can recommend to our patients because we know that it can improve lifestyles”. [Psychiatrist]

7.5.3.8 Suggested Improvements

The study also unearthed a number of suggested improvements for the mobile app. Perhaps unsurprisingly, incorporating support for other social media platforms in the app emerged as a common request amongst participants. One of the young participants said:

“I really like the app [. . .] I think it’s great, but I don’t actually use Twitter. I am on Instagram and use WhatsApp a lot, and my friends are the same. I think if you can have Instagram and WhatsApp, I can see a lot of teenagers using this”. [Young Participant, 15]

As discussed in Chapter 4 (Section 4.5), contrary to initial aims, a number of technical challenges such as API restrictions introduced by Facebook and Instagram and the absence of a WhatsApp API, prevented the integration of BullStop to these platforms. Participants’ desire for the app to integrate with these platforms has, however, provided

additional motivation to explore these limitations and the ways by which a satisfactory compromise might be achieved (see further research discussion in Chapter 9). The other feature popularly requested by participants was to allow the Message Checker to be used across all applications on the phone to serve as an “*offensive content checker*” for the phone as a whole.

In explaining the popularity of icons in computer programs, Hemenway (1982) likened them to geographical symbols on a map. Icons are easier to distinguish between one another than say a set of words and, as such, can convey essential information (such as a warning or command) very quickly. On a mobile phone’s screen, where the display area often needs to be used with parsimony, icons allow mobile applications to signpost users to key areas of the app using the least display area possible. Most of the participants said they understood what the icons used in the app signified from the onset as they have encountered identical icons providing similar functionality in other apps, thus justifying the decision to use the Android Materials Icons in the design (as discussed in Section 6.5). One participant, however, admitted being initially confused with the Sent Messages and Received Messages icons (see FIGURE 7.2) but acclimatised after the “*first few minutes*”. This corroborates the findings of Hemenway (1982), who discovered that while the initial performance of users on icons-only interfaces was poor compared to labels-only and icon-label interfaces, the icon-only interface users very quickly achieved the same level of performance as the other users once they understood what the icons meant. Familiarity with an icon can, therefore, play a part in how users respond to it.

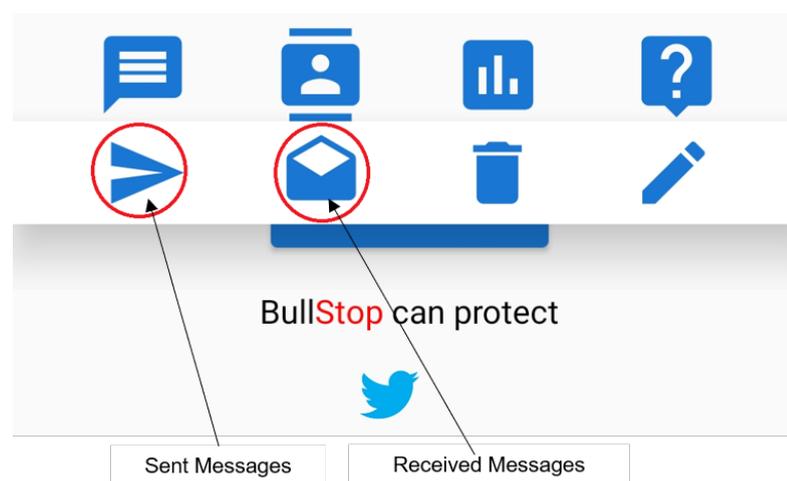


FIGURE 7.2: Sent Messages and Received Messages icons.

7.6 Field-Based Evaluation of the Application's Acceptability

This section reports on the field-based evaluation study of the app that was conducted via the Google Play Store. The study was conducted for fifteen weeks in July - October 2020. Prior to its commencement, the University Ethics Committee reviewed the study protocols, and it was deemed that a separate ethics application was not required (see Appendix E.13).

7.6.1 Study Design

The field-based evaluation study aimed to engage a larger proportion of the target audience to gain an insight into their opinions about the app based on actual app usage in their daily lives. Specifically, it was designed to answer the following research questions:

- 1) Were the intended target audience sufficiently motivated to use the app?;
- 2) What is the perceived usefulness of the app?;
- 3) What is the perceived usability of the app?;
- 4) What are users' favourite aspects of the app?; and
- 5) What areas of the app do users believe require improvement?

Data from the study was gathered via three means: an online questionnaire; application usage data automatically recorded and provided by Google Play Store; and users' in-app actions recorded by the mobile application. The online questionnaire (a copy of which is available in Appendix E.14) was created to capture users' feedback about their experiences using the app. An invitation to complete the questionnaire (in the form of a pop-up window) was displayed to users five days after installation. The questionnaire's prompt window (see FIGURE 7.3) included 'Yes' and 'No' buttons and on tapping 'Yes', the mobile phone's web browser was launched and navigated to the online questionnaire (FIGURE 7.4). If the user selected 'No', then the prompt was displayed as a daily reminder until the questionnaire was completed.

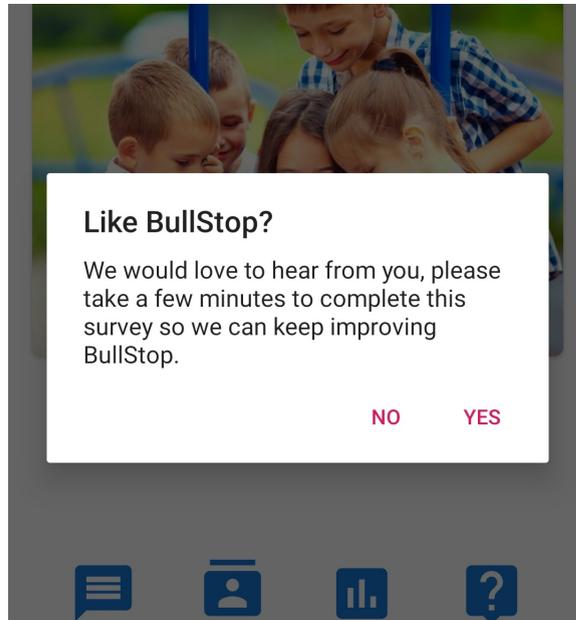


FIGURE 7.3: Sent Messages and Received Messages icons.

BullStop App User Feedback

Consent for Participation in the Study

Thank you for participating in our study. This study is aimed at designing better tools for young people to combat cyberbullying. Your participation is highly appreciated and extremely valuable in evaluating how useful BullStop has been for you and how we can improve it. We aim to make BullStop and similar apps widely available to young people to protect them against cyberbullying.

All questions marked with * are compulsory, and the survey must be completed in one sitting.

By completing this survey, you agree to the statements below. Please take a moment to read them and then click Next when you are ready to proceed. If you do not want to complete the survey, simply close the browser.

1. The information provided by you in this questionnaire will be used for research purposes only.
2. If you choose not to complete this survey, you can continue using the BullStop app. However, you may get reminders within the app to complete the survey.
3. Any information that is obtained in connection with this study and that can be identified with you will remain confidential.
4. You grant permission for the data generated from this survey to be used in the researcher's publications on this topic.
5. Your anonymised data may be used by research teams for future research.
6. SurveyMonkey is acting as a data processor on behalf of the researchers. You can view SurveyMonkey's privacy policy [here](#).

Next

FIGURE 7.4: Online questionnaire welcome page.

App store data was extracted from the Google Play Console (see FIGURE 7.5) which is a feature of the Google Play Store that records information such as user traffic to the app’s listing page, demographic data (e.g., country, age, etc.) and application usage data (e.g., installation, uninstallation, crashes, etc.).

Users’ actions within the application (e.g., screens visited, button taps, navigation patterns, etc.) are recorded by the mobile app using Google Firebase Analytic², an analytic platform for recording user-based events within an Android application (See FIGURE 7.6).

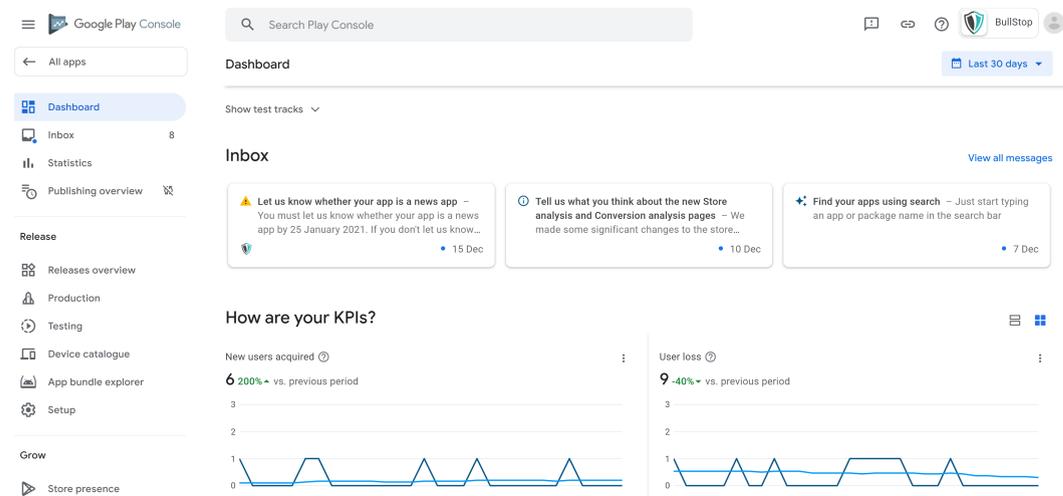


FIGURE 7.5: Google Play Console page for BullStop.

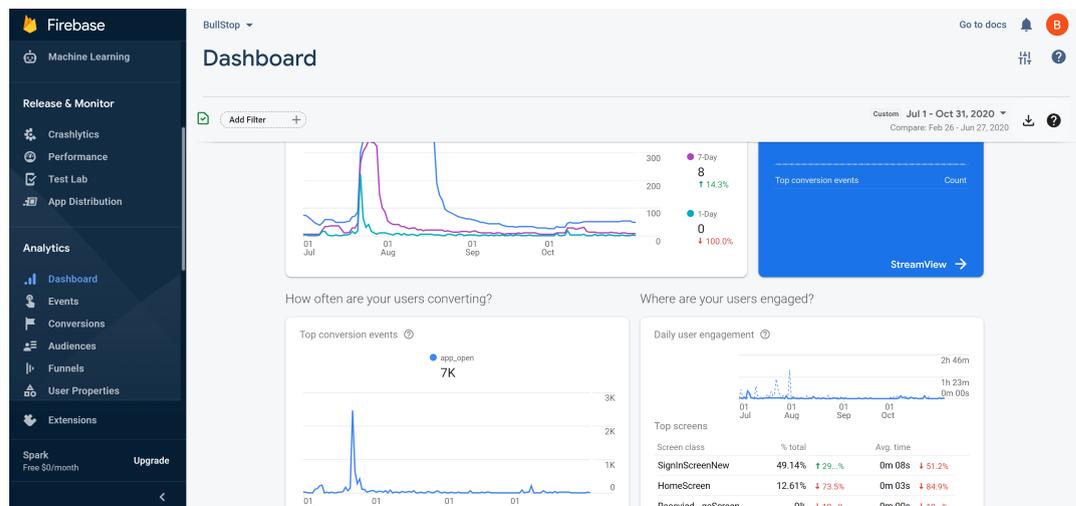


FIGURE 7.6: Firebase Dashboard Play for BullStop.

The mobile app was made freely available on the Google Play Store (see FIGURE 7.7), and the study was publicised via a press release³ issued to media organisations by Aston

²console.firebase.google.com

³aston.ac.uk/latest-news/cyberbullying-shield-app-uses-ai-combat-social-media-trolls

University's Press Office. Additionally, the researcher and his supervisory team shared the press release with personal contacts, including those that participated in the study.

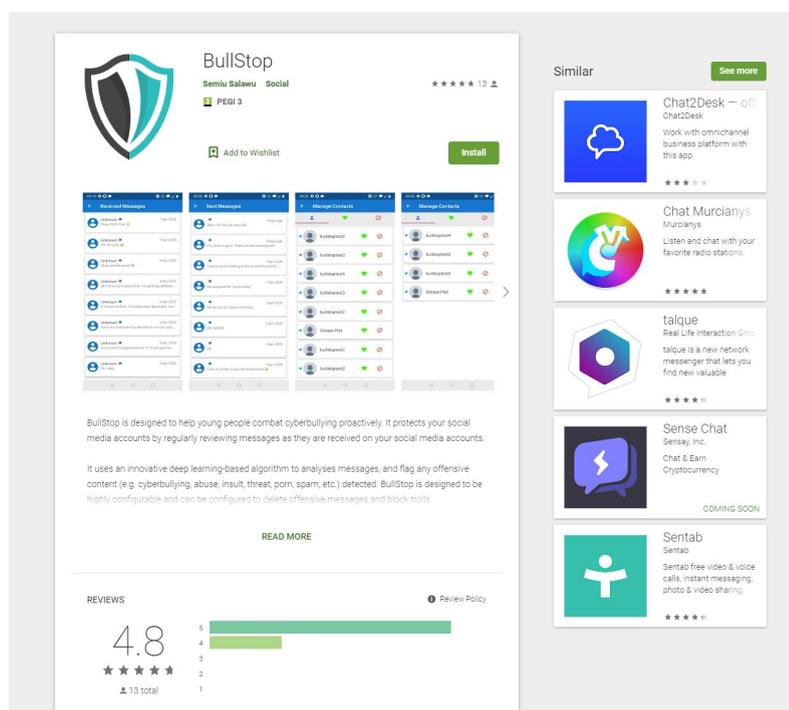
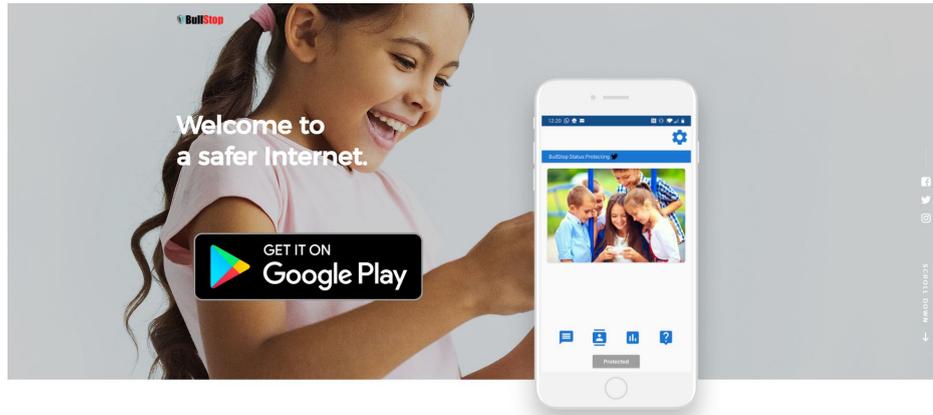


FIGURE 7.7: BullStop listing on the Google Play Store.

7.6.2 Online Presence

Establishing an authentic online presence for the app was identified by the co-designers (during the participatory design phase of the research) as a crucial part of reassuring potential users of the application of its legitimacy (see Section 5.5.5.5). In furtherance of this recommendation, a website (see FIGURE 7.8) and social media pages on Facebook, Instagram and Twitter were created for the mobile app, and a consistent design theme was utilised across all the online channels. As shown in FIGURE 7.9, there was a significant increase in the number of website visitors in the month leading up to the evaluation study (i.e., June 2020) with website traffic peaking in July 2020 when the evaluation study began. This was followed by a sharp decline the month after (i.e., August 2020) and a gradual reduction in traffic in subsequent months during the later periods of the evaluation study. The Google search results for the term 'bullstop', of which the first eight results are related to the app (see FIGURE 7.10), provide an additional attestation of the success of efforts to improve the app's online presence and thus serve as a vehicle for recruiting participants to the study.



About BullStop

BullStop is designed to help young people combat cyberbullying proactively. It protects social media accounts by regularly reviewing messages as they are received. It uses artificial intelligence to analyse messages, and flag offensive content like bullying, abuse, insult, threat, porn, spam and more.

FIGURE 7.8: BullStop app website.

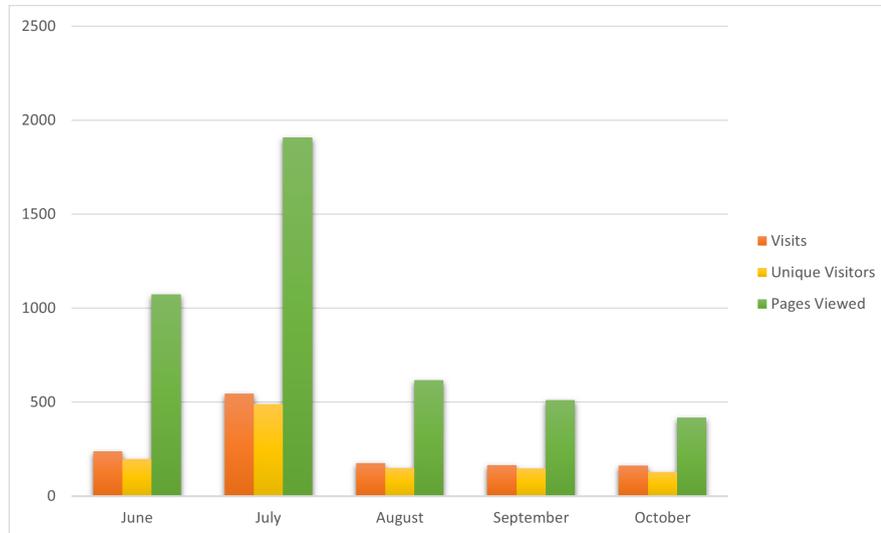


FIGURE 7.9: Traffic to the BullStop website during the evaluation period.

7.6.3 Analysis of Users' Engagement with the Mobile App

As previously mentioned in Section 7.5.1, the application usage statistics available on the Google Play Store and Firebase Analytics were used to understand users' behaviour within the application. FIGURE 7.11 illustrates the number of users gained and lost, active users and visitors to the app's page during the evaluation study. The figures, as well as the definition of the metrics used, were provided by the Google Play Store which defines a *new user* as a uniquely identifiable person (as identified via their Google account) who installs the app for the first time. This metric ensures that a person who installs, uninstalls,

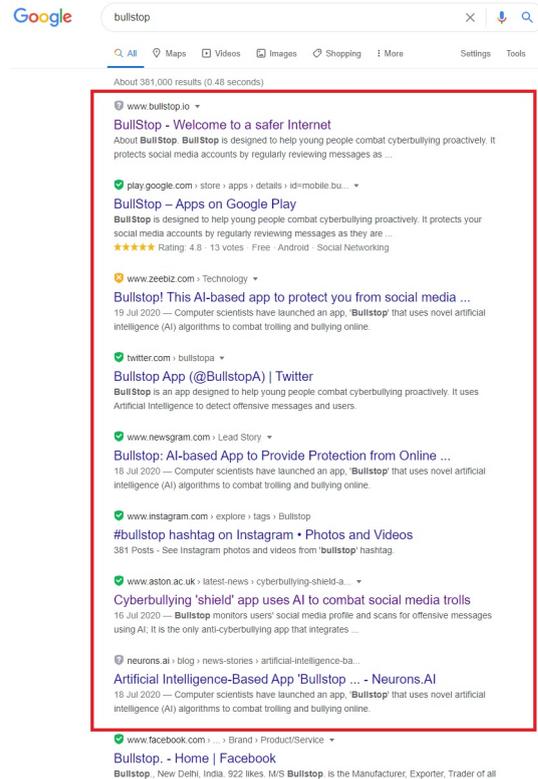


FIGURE 7.10: Google search results for 'bullstop'.

and then re-installs the app is only counted once. Likewise, a user who installs the apps on multiple devices is counted as a single entity. A *lost user* is a user that uninstalls the app from all their devices, and a *visitor* is a person that visits the app's page on the Google Play Store (see FIGURE 7.7). An *active user* is a user who installed the app on one or more devices and has used it in the past 30 days.

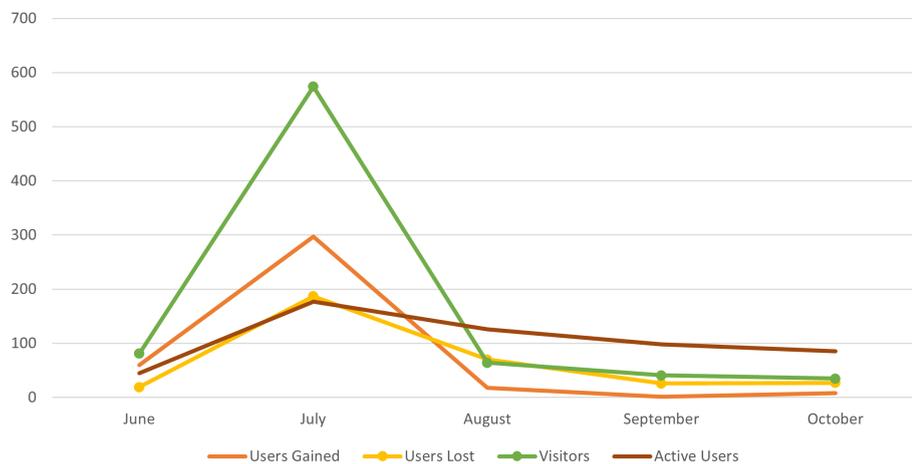


FIGURE 7.11: Total number of active, lost and acquired users and visitors.

During the field-based evaluation study period, there were 795 unique visitors to the app's

Google Play Store page, and 384 of these visits resulted in a new person installing the app (a conversation rate of 48%) with the app being installed 557 times (this figure includes multiple installations by the same user). The impact of the publicity generated by the press release in attracting users is demonstrated by the user acquisition rate before and after the press release as illustrated in FIGURE 7.12 which showed that the majority of users installed the app in the two weeks following the press release.

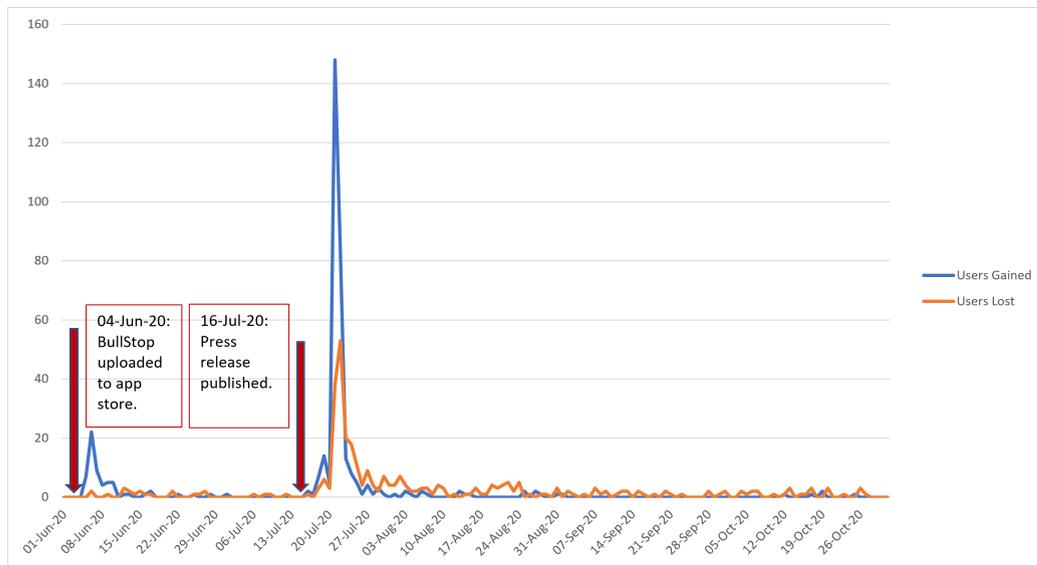


FIGURE 7.12: Users acquisition and loss before and after the press release.

The sheer volume of apps on the app store creates an extremely competitive environment and one in which new mobile apps often struggle to make an impact. As the usage intentions of mobile apps' consumers are heavily influenced by their perception of the app's ability to meet their requirements (Stocchi *et al.*, 2019), a mobile application's app store presence is often the only opportunity available to convince a potential user of its ability to meet their needs. An app's brand then becomes a critical component of its ability to succeed in the app store (Smith and Chen, 2018). This was similarly highlighted by the co-designers' suggestions that the app should have a "cool logo" (see Section 5.5.3.1), "short and catchy name" (see Section 5.5.3.2), and reassure potential users of the app's authenticity (Section 5.5.5.3). The appropriateness of the app's logo and name was validated by the 'lab'-based evaluation study that found participants responded positively to the app's name and logo (see Section 7.4.3.3) and this was further reinforced by the field-based study where 18% of searches for the app used the app's name as the query term (see FIGURE 7.13). This is a remarkable achievement for a new mobile app and implies that despite its infancy, the app's brand was sufficient in attracting users.

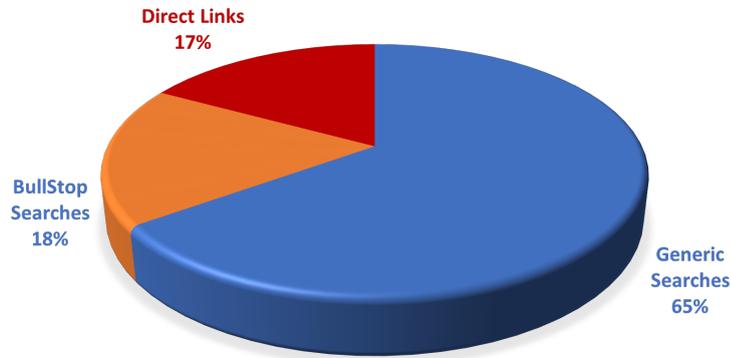


FIGURE 7.13: How the app was discovered on the app store by users.

7.6.4 Analysis of Users' Gender and Age Demography

The gender and age profiles of users (based on their Google account information) suggest that males aged 18 – 24 years make up the largest portion of the app's users. As this is the lowest age range reported by Google Play store, the figure for users younger than 18 years was not available. The male gender's high proportion is not restricted to the 18 – 24 years segment alone. The same pattern is repeated across other age groups, resulting in a predominantly male (80.7%) user base, as illustrated in FIGURE 7.14.

As no consistent interdependency has yet been established between cyberbullying and gender (Calvete *et al.*, 2010; Baldry *et al.*, 2016; Foody *et al.*, 2019), this discovery is not believed to be symptomatic of increased involvement of the male gender in cyberbullying. It may, however, suggest a higher interest level in mobile applications amongst males, as per the study by Seneviratne *et al.* (2015) in which higher male participation rates were discovered when investigating the information revealed about a person by the apps installed on their phones.

Interestingly, the proportion of users younger than 25 years was smaller than expected (35%). A possible explanation of this can be inferred from the age distribution of Twitter users (Statista, 2020b) where under twenty-fives are represented in a similar proportion (i.e., 31%) as shown in FIGURE 7.15. It is therefore understandable that this age group's demand for an app that currently only protects Twitter accounts is lower compared to older age groups. This potentially, however, also makes the app in its current form appealing to older users as they make up the majority of Twitter users.

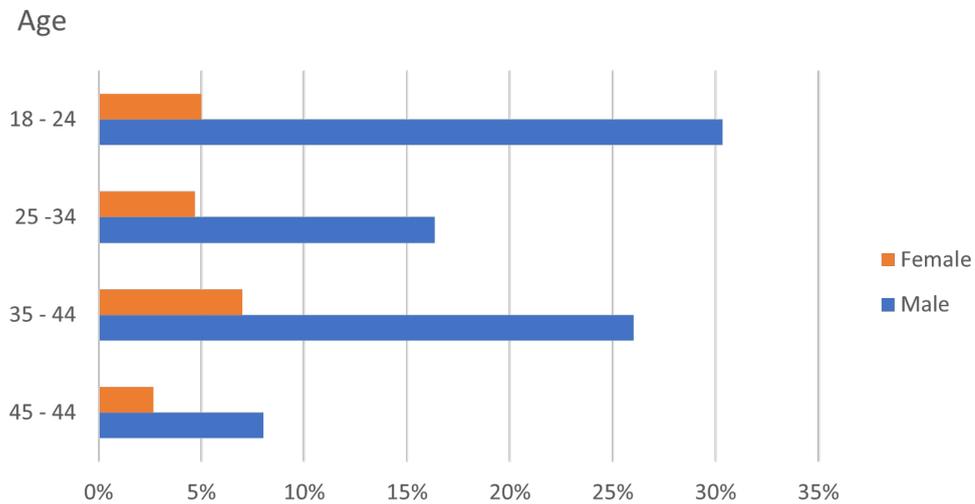


FIGURE 7.14: Age groups and gender of app users.

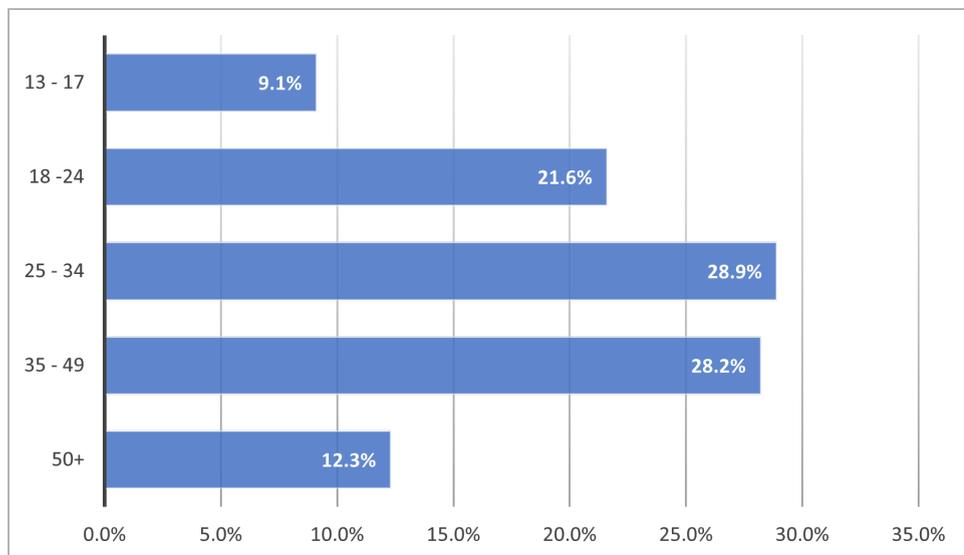


FIGURE 7.15: Age distribution of Twitter users.

7.6.5 Analysis of Application Usage Patterns

The average daily time spent using the app is 2 minutes and 42 seconds and the maximum daily time spent by a single user was 33 minutes and 20 seconds. On average, most time was spent on the Received Messages screen and the least amount of time was spent on the Splash screen, which is displayed when the app is starting on older mobile phones. Of screens that contain interactive UI elements (i.e., text boxes, buttons, links, etc.), the Sign In screen was the screen that users spent the least amount of time on. Again, this is understandable as the Sign In screen is only used to login to the

application after installation and, unless a user explicitly signs out, their session is maintained indefinitely (similar to the behaviour of social media apps like Facebook, Twitter and Instagram).

The users' actions recorded on the Received Messages screens suggest that users typically spend their time on this screen reading the messages, with only about 16% of users selecting or deselecting the offensive label checkboxes to correct the app's predicted labels. In comparison, 57% of the users that visited the Sent Messages screen interacted with the offensive labels' checkboxes and 7% of the users that viewed the Deleted Messages screen tapped on the labels' checkboxes. The implication is that users are more likely to reclassify received and sent messages than deleted messages. This could be because they were sufficiently satisfied with the app's predictions for deleted messages or less invested in improving the app's detection of offensive messages than they were in updating the predicted labels for received and sent messages. This is likely because users feel received and sent messages are more personal and can furnish the app with a more in-depth insight into their communication styles. As such, they may have felt that investing the time to update predictions for these messages would result in more accurate predictions. Furthermore, deleted messages are typically very offensive messages, and users found little reason to dispute the reason for their deletion.

The Message Checker also experienced substantial use as nearly all the users that visited the screen interacted with the UI elements. The Check button (see FIGURE 7.16) was the UI element users interacted with the most on the page, implying that users composed messages and used the real-time offensive content classifier to predict offensive labels for their messages. The Message Checker's popularity amongst users corroborates a similar finding from the 'lab'-based evaluation study and supports the findings of Dinakar *et al.* (2012) that the use of reflective interfaces encourages more empathic behaviour in cyberbullying situations amongst users, thus reducing their proclivity to offend.

Table 7.3 presents a list of the top ten screens based on the average daily time spent there by users. The second-highest average time was spent on the Detection Settings screen, and most of this time was spent by users adjusting the Deletion Threshold slider control to different values. The inference from the recorded events around this activity suggests that users adjusted their Deletion Threshold setting and then visited the Deleted Messages screen moments later. This is likely so that they can access the new settings'

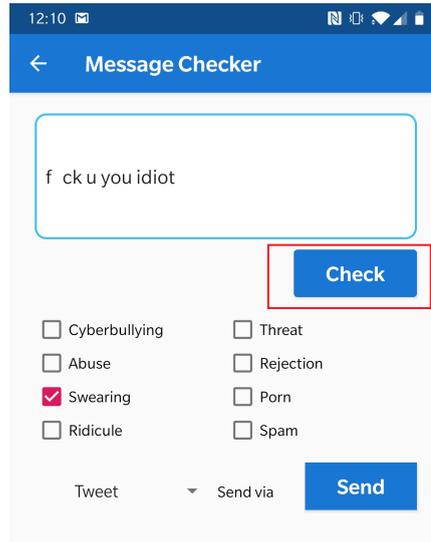


FIGURE 7.16: The Message Checker screen's check button.

Screen	Average Time Spent by Users (seconds)
Received Messages	26
Detection Settings	19
Sent Messages	17
Message Checker	16
Sign-Up	13
Tour	8
Reports	7
Home	5
Anti-Bullying Helplines	4
Forgot Password	4

TABLE 7.3: Top Ten screens users spent the most time on.

impact, perhaps by sending themselves messages with varying levels of offensiveness. As is to be expected, the Sign-Up and Home screens were also frequently accessed by users. The Sign-Up screen is used to create a BullStop account after installation of the app, and as this is a prerequisite to using the app, all users will visit this page. Likewise, the Home screen, which as the application's main interface, is used to access other areas of the app.

While the app's Tour, Anti-Bullying Helplines and Report screens could be considered secondary to the app's core purpose of online abuse detection, their inclusion in the top ten screens list indicates more user activities on these screens than anticipated. The average time spent by most users on the Tour screens is less than was envisaged it would take to read the tour's content, which suggests that many users visited the screens out

of curiosity rather than to understand how to use the app in detail. This could imply that users found the app easy enough to use without the need to consult the tour screens. Three calls were initiated from the Anti-Bullying Helplines screen to two of the listed charity organisations, namely Bullying UK and SupportLine . There is no more information beyond this regarding these calls as the information collected by the app does not extend beyond users' interaction with UI elements. The Reports screen provides a simple count of the number of messages deleted and contacts blocked and requires little user interaction, yet surprisingly it accounts for seven seconds of the time spent by users on the app.

The Forgot Password screen was another unexpected inclusion in the list, and this can be attributed to users using the password reset facility to change their BullStop accounts' passwords. Interestingly, the contacts screen which generated many discussions during the design sessions was not amongst the ten screens. This could be because of the way people typically use Twitter. It is primarily used as a means of keeping abreast of new developments in areas of interest and less as a means to stay in touch with friends and acquaintances in the same manner as Facebook and Instagram.

7.6.6 Findings and Discussion

Subjective feedback from the application users was gathered via an online questionnaire accessed from within the app through a web link. This link contained a unique code generated for each user that was used to monitor questionnaire completion for each user (so that the questionnaire reminders could be disabled). The questionnaire response rate was approximately 11%, representing 43 out of the 384 app users. The questionnaire responses and the application usage data recorded were used to answer the research questions posed as investigative directions for the field-based evaluation study. These are discussed in the following subsections.

7.6.6.1 App Usage by Young People

Of the forty-three completed questionnaire responses received, users aged 13 – 15 years and 16 – 18 years represented 35% and 21%, respectively. The target audience, therefore, represented more than half of the survey respondents (56%) and over a third of all users. Additionally, the gender distribution for the questionnaire respondents was more evenly

distributed compared to the gender distribution for all users. There were 19 female, and 24 male respondents with an age distribution as shown in FIGURE 7.17 which equates to a higher proportion of overall female users completing the questionnaire (about 26%) than male users (8%).

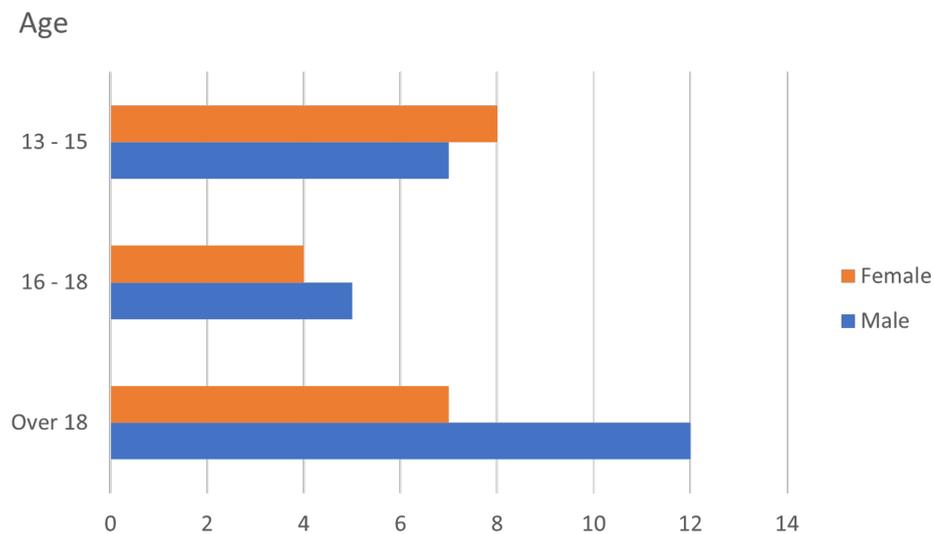


FIGURE 7.17: Age groups and gender of respondents.

About 30% of respondents had experienced online abuse, and 19% had bullied others online before. The majority of users that had been involved in cyberbullying as either victims or abusers were 18 years and under and many of these incidents had occurred as recently as within the preceding seven months. These findings suggest that the primary target audience was sufficiently motivated to use the app and report their observations via the questionnaire thus validating the results' representativeness.

Respondents' use of social media (see FIGURE 7.18) is similar to earlier findings from the pre-interview survey. As reported in Chapter 5 (Section 5.3.1), Instagram, Facebook, WhatsApp, Twitter and Snapchat remained the most popular social media platforms used and 98% of all respondents noted accessing these platforms a least a few times a day.

Word of mouth and the app store emerged as the two primary means via which respondents found out about the app, followed by online articles about the app and web searches (see FIGURE 7.19). As these are not mutually exclusive, it is more likely that a combination of these contributed to raising awareness of the app amongst potential users.

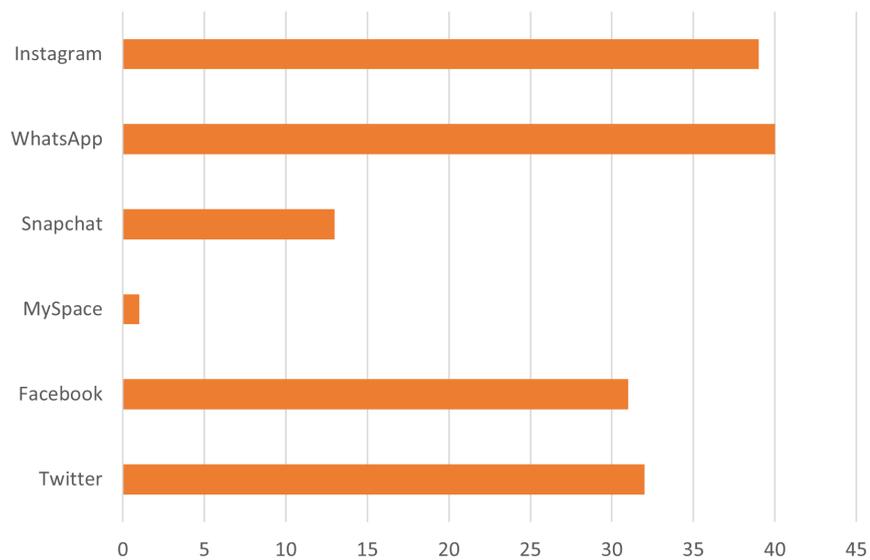


FIGURE 7.18: Social media platforms usage amongst BullStop users.

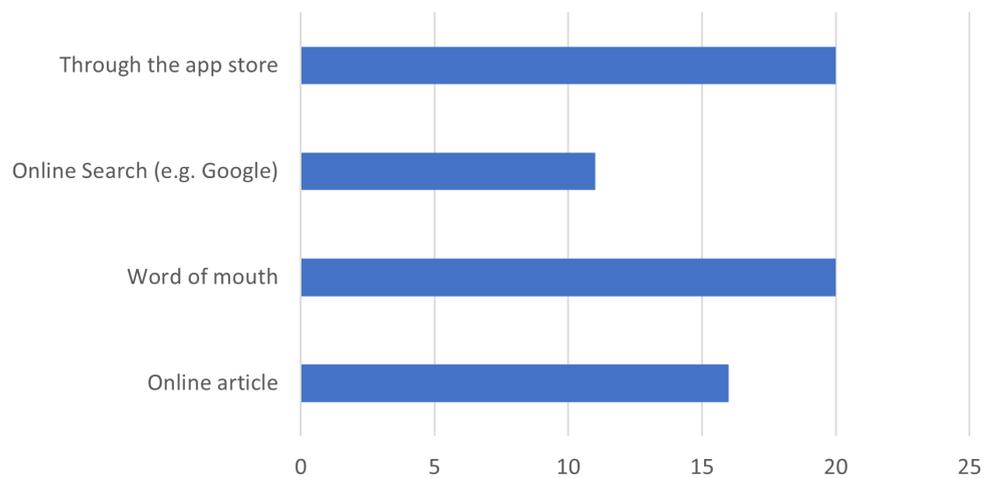


FIGURE 7.19: How BullStop users found out about the app.

7.6.6.2 Perceived Usefulness of the App

In seeking to understand users' impressions of the app, they were asked the question "Do you think the app is a good idea?" All respondents answered in the affirmative and provided the following example comments as to why they thought so:

"Because people can get hurt from cyberbullying and it can cause mental issues".

"It is very educative and can help teenagers to fight bullies".

"It can help alot of people with cyberbullies".

"Bullying is a big problem and the social media companies are not doing enough about it".

"It's good to educate people about cyberbullying".

"I think 'it's a very good idea and 'I'm happy I found it".

"Because it protects people from cyberbullying".

"It's protects from abusive words on social media".

"It is a good idea because it will help prevent people using it from being cyberbullied as it blocks such messages from coming through".

"In this era of social media and the influence it has on people, cyberbullying has become a thing in many homes and has affected individuals mental health".

"Cyberbullying is real, and have a significant impact on victims. It is one of the types of abuse that has deep underlying damage which takes time to surface and as such needs to be taken seriously".

"Because cyberbullying is a big problem".

"Good app for stop cyberbullying".

The comments and responses indicate respondents had a very positive disposition to the app and in a following question that asked *"Do you think the app would be useful to people of your age?"*, all but one respondent responded with a "Yes". Further analysis revealed that the user who answered "No" was aged over eighteen years (according to his/her self-reported age) and as the reason provided for this response was: *"I think it is more appropriate for teenagers as I can handle trolls by myself"*; this is, therefore, actually considered a positive response as it still indicates the app's appropriateness for the primary target audience.

The app's ability to automatically detect and remove offensive messages is fundamental to its usefulness as a cyberbullying detection and prevention tool. Users were therefore questioned about their perception of the app's performance in this regard. Twelve users indicated that the app deleted offensive messages for them. Five of these users rated the app's ability in this regard as *"very good"*, six rated it *"good"*, and one user felt it was *"average"*. No respondent thought it was *"bad"* or *"very bad"*. The maximum number of offensive messages deleted by the app for a single user was thirteen, and five users said they updated the app's prediction for some of their deleted messages. These updates were to further qualify the deleted messages by associating additional labels with the messages;

as such, no inoffensive message was mistakenly deleted by the app during the evaluation study.

Eleven users said they updated the predictions for their received messages by associating other applicable offensive labels to the messages, while eight users did the same for their sent messages. The higher number of users who reported updating the predicted labels for received and sent messages compared to those that did the same for deleted messages corroborates the increased user events recorded by the app for the offensive label checkboxes in both Received and Sent Messages screens compared to the Deleted Messages screen (see Section 7.5.5). A possible explanation for this could be that users updated the predictions for their sent and received messages more than the deleted messages because they believe the app could learn more about how they communicate from sent and received messages.

As reported in Chapter 4, correctly identifying inoffensive messages so that they are not mistakenly deleted is equally as important (or more, for some interviewees) as detecting offensive messages. Users' opinions about the app's performance in this task were, therefore, crucial in reflecting their perceptions of the usefulness of the app. Most users (81%) rated the app as "good" or "very good" on this task, and the lowest rating received for the app on this was "average".

Finally, when asked to provide an overall rating for the app, 88% of users rated the app as "very good" or "good" and the remaining users awarded an "average" rating (see FIGURE 7.20). Additionally, 63% of respondents said they would continue using the app, and for the respondents who indicated they would discontinue use, not being able to use the app with other online social networks was the main reason cited for their decision; given this limitation was imposed on rather than within the immediate control of the research, it is not considered a negative reflection on the core principles the app is trying to support. Findings in relation to the research question "*What is the perceived usefulness of the app?*" would therefore indicate that it was generally perceived to be useful, at least within scope.

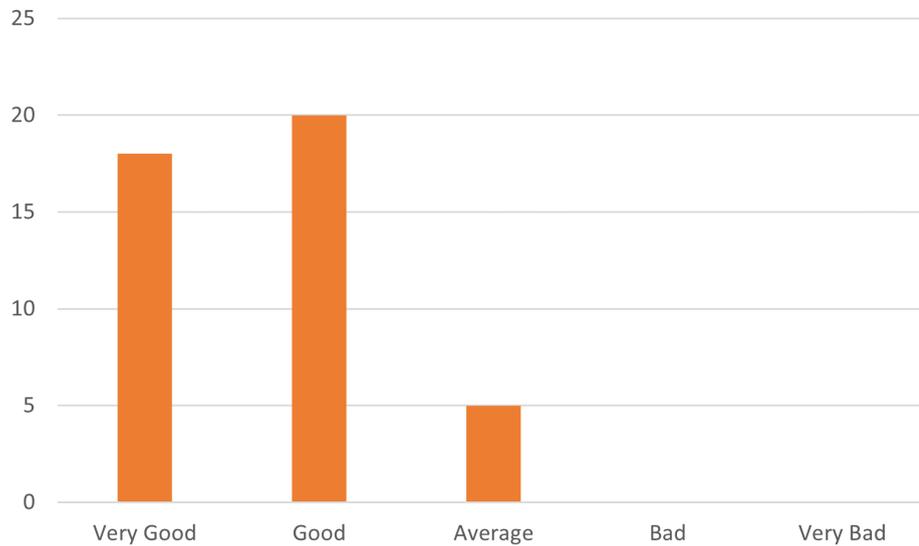


FIGURE 7.20: Users overall rating for the app.

7.6.6.3 Perceived Usability of the App

86% of respondents said they found the app “easy” or “very easy” to use, while 12% said it was “neither easy nor difficult”. Only one respondent said (s)he found the app “difficult” to use. The same individual confirmed that (s)he took the app tour, which (s)he found useful, and would have struggled to use the app without the tour. Three other respondents also confirmed that the tour was vital in educating them about using the app. For these individuals, the app was “easy” or “neither easy nor difficult” to use. The majority of respondents, therefore, found the app easy or very easy to use and, for those who did not, the app tour served its purpose by bridging the knowledge gap for such users.

When asked the question: “Do you feel that the app was well designed for people your age?”, all respondents replied “Yes” – an affirmative response – and, in so doing, validated the PD-led approach used when designing the mobile app. It is particularly encouraging that the mobile app meets respondents’ design expectations considering their diverse age groups.

While the majority of respondents were satisfied with the icons used, three complained that not all of the app’s icons were self-explanatory. Since a participant in the ‘lab’-based evaluation study had expressed a similar sentiment, it would appear that even the use of standard and recommended icons such as the Android Material Icons does not guarantee total acceptance amongst users, highlighting a potential area for future work in

improving user satisfaction with the app icons, perhaps in the form of additional end-user engagement to agree on a set of icons.

7.6.6.4 Perceived Responsiveness of the App

As mentioned in Section 7.3, users' perception of the application's responsiveness is a critical factor that is perhaps more important than the actual average response time recorded in the responsiveness evaluation experiments (see Section 7.2). The 'lab'-based study had already provided some indication of users' positive opinions on the app's responsiveness (see Section 7.4.2.4). The field-based study was therefore used to expand on this and garner more feedback regarding this from the beta users. Respondents were asked to provide a rating for the system's responsiveness, and all the surveyed users rated the app as "very good" (the majority) or "good" (FIGURE 7.21). Considered in tandem with the average response time of about 1 second achieved in the computer-based evaluation experiments conducted, the application can be judged to be highly responsive based on these findings.

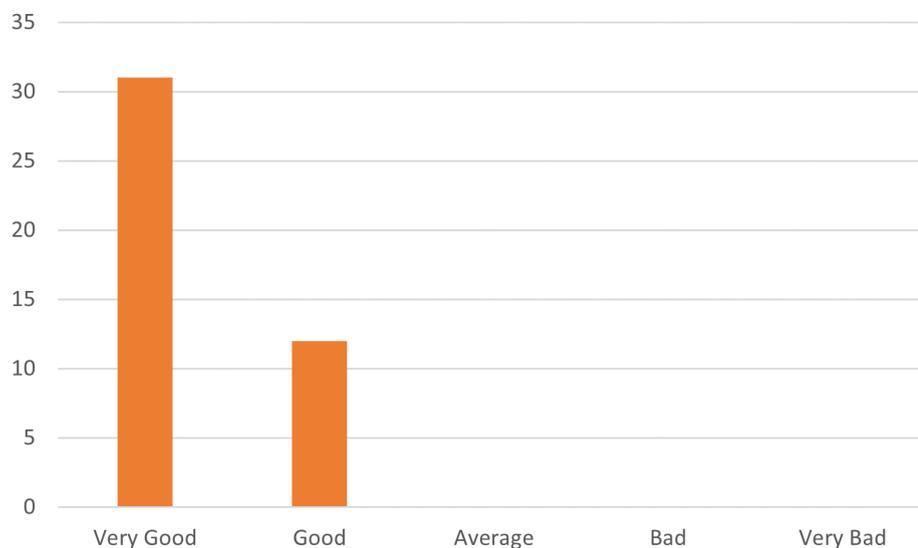


FIGURE 7.21: Sent Users responsiveness rating for the app.

7.6.6.5 Users' Favourite Aspects of the App

To gain insight into users' preference for the app's various components, feedback was solicited from respondents on the features of the app they liked the most. Their

responses provided a rich set of data, highlighting the popularity of app features amongst respondents'. They liked the app's design and ease of use. The Message Checker, Contacts and Anti-Bullying Helplines emerged as favourites for many respondents, many of whom also appreciated that the app did not only flag offensive content in received messages but did so for messages sent by the users as well (another reflective feature of the app). Respondents liked that (a) the app runs in the background without draining the phone's resources, (b) being able to update the predicted labels, (c) the app's integration with Twitter, and (d) its ability to automatically block abusive contacts. Sample comments relating to the respondents' favourite aspects of the app include:

"I like the message checker, that is quite good".

"The helplines are a good idea".

"Being able to tick the boxes so I can improve the app".

"That it can block bullies automatically"

"I like that it can bring in all my Twitter followers with their pictures. That's pretty cool".

"Flagging my comments as being offensive. This will further entrench the culture of thinking before posting on social media platforms".

"The idea that it works in the background with minimal data usage and reviewing battery usage, that was minimal too".

"Its ability to sync easily with the social media apps".

"The AI".

"Simplicity to install and use".

"I like the design; it's very nice".

7.6.6.6 Suggested Areas for Improvement

While the study discovered some variety in respondents' favourite aspects of the app, there was considerable consensus regarding the area most in need of improvement. The app feature most often suggested as an improvement was the need to make other social media platforms available within the app, as illustrated below:

"Not everyone uses Twitter".

"I would like more social media networks".

“If it can work with WhatsApp.”

“Please add Instagram, YouTube and TikTok”.

“I don’t use Twitter”.

“It would be nice if [the app] works with other social media networks”.

As previously identified, a small number of respondents expressed some confusion about some of the app icons and this was found to be the second area identified for improvement, as typified by the following comments:

“Some icons were confusing, but I like the drop menu”.

“Not all the icons are not self explanatory”.

7.7 Summary

This chapter details the evaluation studies (and associated results) conducted as part of this PhD research to evaluate the scalability, responsiveness, acceptability, and usability of BullStop. Technical experiments conducted to assess the system’s scalability and responsiveness demonstrated that the system adequately addressed these concerns. In addition to the computer-based evaluation studies, two human-focused evaluation studies were conducted to assess users’ perceptions of the application’s user interface design, performance, perceived ease of use, responsiveness and brand. The first human-focused evaluation study was ‘lab’-based and comprised a series of exploratory sessions with selected participants representing stakeholders, followed by in-depth discussions of their experience using the app for a limited period. The second (field-based) evaluation study was expanded to the general public at large and used mechanisms embedded in the application to collate information about how people discovered and used the app, and an online questionnaire to gather users’ opinions about the app.

The most frequently requested extension to the app’s functionalities across both evaluation studies was expanding the app’s integration to other online social networks, aside from Twitter. The Message Checker emerged as the favourite component of the app for most users, with the ‘lab’-based evaluation participants requesting for this feature to be expanded such that it can serve as an “offensive content checker” for all applications installed on the phone. As the Message Checker is an educational tool that allows users

to reflect on how they communicate with others, its popularity among users suggests that using reflective interfaces to reduce cyberbullying offending is a viable strategy.

While users rated the app highly for its ability to correctly detect and delete offensive messages, its performance in identifying instances of sarcasm, social exclusion, and threat within these messages was identified as an area for future improvement. The app's performance in this regard is heavily influenced by the low number of examples of these types of offensive content present in the training data used to train the classifier; as such, increasing their distribution in the training data is an area for future research. Furthermore, the subjective nature of cyberbullying necessitates additional consideration of how its detection is framed. Thus the app provides the means for users to update the offensive labels predicted for messages, and over time this will improve the app's ability to identify different forms of online abuse.

Users also appreciated the app's responsiveness, a feature that can be directly attributed to the app's use of a distributed microservices architecture that makes use of cloud-based technologies to offload computing-intensive operations, such as offensive content detection, to the cloud (as discussed in Chapter 6). Finally, brand appeal and positive publicity have been shown to be a key component in attracting potential users to the mobile application.

The use of UCD and PD approaches in designing the mobile application led to a comprehensive understanding of the stakeholders' requirements for the application and provided valuable insight from the unique perspectives of young people who have experienced cyberbullying as offenders, victims, and/or bystanders. This knowledge has proven invaluable in the design of the app, in informing the functionalities to include, and in determining how these should be implemented. The proportion of young (36% for 'lab'-based evaluations and 56% for the field-based study) and adult users that participated in the evaluation studies ensured inclusion of both adult and young stakeholders' opinions of the mobile app. The overwhelmingly positive feedback received from the evaluation studies has demonstrated the merits of both user-centred and participatory design approaches in designing applications, not just for young people, but any audience, and will hopefully encourage other researchers to adopt similar techniques in the future.

Chapter 8: Reflections on the Design Approach

8.1 Introduction

The adoption of user-centred design (UCD) and participatory design (PD) approaches in the design of BullStop was born out of a desire to address the paucity of research into the development of cyberbullying prevention tools the focus of which extend beyond the implementation of state-of-the-art algorithms to improving the usability and fitness for purpose of the cyberbullying prevention tools themselves. The methods used have been instrumental in understanding and subsequent implementation of the stakeholders' collective desires in the developed mobile app. While the research has derived and evidenced tangible benefits from application of these methods in the form of users' high levels of perceived usefulness and usability of BullStop, researchers (especially those in the computing field) may still be wary of adopting UCD and PD techniques on account of the high level of end-user engagement required, especially when dealing with sensitive topics such as cyberbullying amongst young people. This chapter presents the learning outcomes realised during the process and reflects on the researcher's observations and experience of applying these methods, as well as the experience of the various participants involved at different stages of the process, in the hope that it will inspire (and help) other researchers to follow suit.

8.2 Learning Outcomes

The transcripts from the various recordings of participants' engagement activities yielded a rich collection of qualitative data that was repeatedly referenced throughout the app's design and development process. While the participants' views during the discussions directly shaped the requirements for the mobile app, their remarks about the process

itself provided an insight into their opinions about the knowledge elicitation activities. The learning outcomes reported here can be considered part reflections and part recommendations expanding on existing recommendations for implementing participatory design.

8.2.1 Reflection #1: Inclusion and Representation are Vital

While cyberbullying is often viewed as a concern for young people, parents, and teachers (Dehue *et al.*, 2008; Eden *et al.*, 2013; Makri-Botsari and Karagianni, 2014; Łukasz and Anna, 2019), its negative consequences on society extend beyond the impacts felt by these groups. It was, therefore, crucial to fully explore all the relevant groups affected or interested in cyberbullying and its prevention to ensure adequate representation of the different views on the subject. In identifying this study's stakeholders, the process started with the aforementioned groups (young people, parents, and teachers) as the initial subset of stakeholders and, via initial conversations with these groups, other potential stakeholder groups such as clinicians and law enforcement were identified. Further discussions with the additional stakeholder groups revealed sub-divisions within some of the groups, providing opportunities for the study to explore other perspectives and thus further enriching the acquired data and the findings derived from them. For example, this study's initial clinician recruits were mental health professionals but, in exploring how these professionals engage with cyberbullying victims, the role played by General Practitioners (GPs) as the initial point of contact for cyberbullying victims was realised. The direct consequence of this was an expansion of the focus groups to include a GP, thus enriching the ensuing discussions with an additional and unique perspective. Additionally, in engaging with the different stakeholder groups to understand their expectations of the proposed mobile application, the study acquired a varied set of requirements for the application that contributed to its appeal to not only the primary target audience (i.e., young people) but also to age groups older than the primary target audience as evidenced by the substantial proportion of adults who were motivated to install and use the app.

Based on the practical experience of this research, it is therefore recommended that when engaging in UCD/PD, it is crucial to fully explore who the interested parties might be for the topics being investigated. Researchers must be open to seeking out different

perspectives by engaging with a broad cross-section of people whose lives may be impacted by the researched topic. Furthermore, it should be recognised that sub-divisions may exist within stakeholders' groups and identifying these can be the key to discovering additional perspectives.

8.2.2 Reflection #2: Build Rapport with Participants from the Onset Via Frequent Communication

MacDougall and Fudge (2001) proposed a prepare, contact, and follow-up process when recruiting participants for interviews and focus groups, and this study's approach bears similarities to their proposition. The recruitment process started by identifying personal contacts who belonged to potential stakeholder groups and initiating contact with them. For other stakeholder types for which the researcher did not have personal contacts (e.g., law enforcement), potential recruitment sources were identified and engaged (e.g., the College of Policing in Coventry). This extended engagement approach embodies the recommendation by Lumsden *et al.* (2017) to avail opportunities to work with existing support organisations as part of the recruitment process. Once prospective participants had been identified, regular communication was maintained up to and following the UCD/PD sessions. This contact was not just in the form of reminders about the scheduled sessions, but continued throughout the different research stages and included regular updates about the research progress to participants. The importance of keeping participants informed about the process was reinforced by Kensing and Blomberg (1998) Kensing (1998) who identified access to relevant information as a key PD requirement. Similarly, Lumsden *et al.* (2017) advocated communicating the nature of their involvement to participants and, by ensuring that participants remained well informed through the process, this study encouraged 'buy-in' from participants. One of the focus group participants noted:

“ I wanted to drop off [. . .], but your updates were very interesting. I kind of wanted to know how things end.”

This 'buy-in' ensured good rapport with the participants and subsequently created a very interactive environment resulting in in-depth discussions and valuable insight.

Based on practical experience, the importance of regular communication with participants from initial contact through the different stages of the process is highlighted as key to keeping participants engaged and ensuring a sociable environment conducive for collaborative work. Ensuring a friendly atmosphere was identified by Lumsden *et al.* (2017) as a crucial component of PD and the above recommendation extends their guideline by proposing frequent communication with participants as a means to provide this.

8.2.3 Reflection #3: When Soliciting Knowledge from Participants, Be Prepared to Introduce Procedural Flexibility to Progress

The knowledge gathering phase involved the use of both qualitative and quantitative methods, where methods were selected to be appropriate to the profile of participants, their availabilities, and the aims of specific engagements. In the focus groups, the adult participants reflected on each other's views and exposed new areas of inquiry for the researcher to explore in the subsequent sessions. By interviewing adolescents individually, they were able to discuss past experiences (such as them bullying other people) in a way that may not have been possible within a group. The pre-interview survey allowed the study to acquire relevant background information about interviewees and create a shortlist of participants for the interviews. Finally, the PD sessions were integral to bringing together all the various forms of acquired knowledge into the design of the app's prototype.

Challenges encountered at different stages of the research required introducing some procedural flexibility in the approach. Examples of these include the difficulties experienced in recruiting secondary school students as participants which necessitated using first-year university undergraduates instead. Additionally, some 'lab'-based evaluation sessions were held at cafés close to some of the professionals' workplaces to resolve their limited availability. Likewise, the law enforcement officer was offered the option of video-calling into one of the focus groups when he initially thought he would not be able to attend in person but this later proved not to be necessary.

The use of Pair Design and Layered Elaboration to manage conflicting ideas during the PD sessions and motivate co-designers towards a common design vision is another example of a situation where it became necessary to introduce procedural flexibility to advance the study. While these techniques are not PD methods per se, their use at a critical moment was crucial to progress the design. Researchers should, therefore be aware that it may become necessary to amend their approach to progress their research. Crucially, this reinforces Hakobyan *et al.* (2014) acknowledgement of the methodical concessions required to advance the work in their PD approach to developing assistive technology for vision-impaired older adults.

8.2.4 Reflection #4: Maximise Opportunities to Engage with Participants

Co-designers are the essence of participatory design, yet can simultaneously represent some of the approach's strengths and weaknesses. While their inputs enrich the design process with valuable perspectives that would otherwise be missing, poorly engaged participants can delay or even derail the process (Massimi and Baecker, 2006). This heavy reliance on well-engaged participants was an issue that the researcher had to grapple with on several occasions during the research programme. It is also likely a contributor to developers' wariness for the process and the restricted use of PD within the software industry (Kensing and Blomberg, 1998). Researchers should treat participants as scarce resources available only for a small amount of time.

Related to Reflection #3's recommendation for researchers to be prepared to incorporate flexibility in their adoption of PD is the need to understand how best to maximise the engagement with participants while considering individual circumstances. For example, in dividing the focus group participants into the two groups as performed in this research, participants with limited availability were placed in the same group and attended two sessions instead of the three held for the second group. The second session for the limited availability group was extended to accommodate discussions on the topics planned for the third session. When reflecting on the PD approach with the co-designers during the final PD session, the group indicated that the non-engagement on their part that impacted aspects of the research could potentially have been mitigated by organising the three planned PD sessions as a day-long event to progress the design through

conceptualisation to low-fidelity prototyping and on to the final high-fidelity prototype on the same day.

Of course, the applicability of such proposals depends on the type of participants involved. While the co-designers in this case were young adults who may have been able to cope with and accommodate a day-long design session, the same cannot be said of other demographics. In seeking to maximise engagement with participants, researchers must give due consideration to participants' welfare and circumstances.

8.2.5 Reflection #5: Participants Empowerment Encourages Active Participation

It was discovered that being involved in the study was an empowering experience for participants, and this, in turn, encouraged more active contributions from participants as exemplified by the remark below:

"I feel like I'm an expert on cyberbullying now [...] I enjoyed the discussions, so I wanted more."

Some of the co-designers also noted that:

"I got a confidence boost from doing the app's design. I didn't think it was something I was capable of".

"I will definitely spend more time thinking about my app ideas and maybe use the proto software to sketch them out".

Researchers can, therefore improve participants' engagement with the study by incorporating the means for them to acquire skills or knowledge that they may find useful outside of the study.

8.3 Summary

As previously noted, PD has a heavy reliance on highly motivated and well-engaged participants, and researchers should start the process by nurturing a shared vision with co-designers. To gain the full benefits of PD, researchers must be prepared to embrace

its democratic underpinnings wholeheartedly, while simultaneously making co-designers aware of the technological impacts and constraints of design proposals and serving as the 'voice of reason' to mediate conflicting co-designers' visions. Additionally, co-designers must be suitably empowered as part of the decision-making process, and their contributions valued even if they represent different positions to that of researchers. The recruitment of participants should seek to address all relevant perspectives equally to ensure adequate representation of all stakeholders.

This chapter presents key learnings from the research programme's use of a participatory design approach to the development of the BullStop mobile application. These reflections are provided as suggestions to assist future researchers to overcome some of the challenges inherent in adopting a PD approach when designing technological solutions for young people. It is hoped that these can signpost potential problems before they occur and help researchers successfully implement PD.

Chapter 9: Conclusions, Contribution to Knowledge, and Further Research

9.1 Thesis Conclusion

This dissertation has detailed research conducted to design, develop, and evaluate a novel mobile application to detect and prevent and mitigate cyberbullying and online abuse on social media. Primarily targetted at young people, the innovative application has been co-designed with stakeholders to ensure it will appeal to users of all ages. The research explored stakeholders' opinions on cyberbullying and its prevention and the specific ways that the mobile application can assist and protect online abuse victims. The research presented in this dissertation was conducted via five key phases: (1) the creation of a novel large-scale cyberbullying dataset to capture various forms of online abuse and facilitate the training of ML models which serve as cyberbullying detection classifiers; (2) discovery of end-users' and other stakeholders' expectations for the proposed cyberbullying prevention mobile application; (3) the use of participatory design techniques (novel within this field) to create a high-fidelity prototype of the proposed tool; (4) the development of the innovative BullStop mobile application and the extensible and scalable framework that it uses to address the challenges of scalability, responsiveness, and obsolescence; and (5) evaluation of the BullStop application to validate its capacity to address these challenges, especially from the perspective of end-users (such evaluation being exceptionally rare in this field).

Following a review of cyberbullying detection approaches that use various machine learning techniques, and the datasets used to facilitate the implementation of these techniques, this research programme developed a novel, large-scale cyberbullying dataset to address some of the observed shortcomings of existing datasets for use in this

research domain, such as low proportion of offensive content and inadequate consideration for less frequently encountered forms of online abuse. A new labelled dataset comprising 62,587 tweets was created and used to train different deep-learning and traditional ML models. A range of experiments conducted using the dataset identified RoBERTa as the best performing model and demonstrated the dataset's cross-domain applicability. The collection strategy for creating the dataset was designed to target cyberbullying and offensive tweets and ensure that these tweets constitute the majority class. As the occurrence of cyberbullying documents is naturally low, classifiers trained on this dataset can benefit from a high concentration of cyberbullying and offensive documents without the need for oversampling techniques. The imbalanced nature of the dataset was not found to affect the abilities of models trained with it to learn both offensive and non-offensive content. Using the dataset, therefore, saves researchers the effort of simulating real-world distributions of offensive content and implementing oversampling techniques to improve offensive content proportions within existing datasets.

In furtherance of the research programme's aim of creating a viable and impactful tool for cyberbullying prevention, a series of UCD-based activities were conducted in the second phase to solicit stakeholders' perspectives on cyberbullying prevention and gain insight into their vision for actualising these in the proposed mobile application. Analysis of the study findings revealed some key strategies to aid cyberbullying detection and prevention, particularly in educational settings. These include promoting an environment where young people can safely report cyberbullying incidents to school authorities without fear of reprisals from the perpetrators, and fostering positive relationships with peers since cyberbullying victims were found more likely to confide in their friends about experienced online abuse than they would confide in authoritative adults in their lives. In addition, the availability and presentation of relevant advice on cyberbullying prevention were highlighted by participants as key issues negatively impacting the success of existing cyberbullying prevention strategies. Adult participants indicated a preference for cyberbullying prevention advice to be presented in an easily digestible format like a 'cheat sheet' while young people emphasised the importance of cyberbullying advice to be non-patronising and created by people with a good understanding of how cyberbullying is perpetrated amongst young people and its effects on them. Focus groups and interviews were conducted that facilitated the identification of a core set of reflective and punitive features desired in the mobile application by stakeholders. Reflective elements that

encourage positive online behaviour amongst young people emerged as the most popular feature for all stakeholder groups, and the use of monitoring facilities to 'spy' on young people was identified as the least popular feature across the studies. These then formed the basis of the next (and third) phase of the conducted research, which employed participatory design techniques to work collaboratively with young people as co-designers to create high-fidelity prototype designs for the mobile application

The outcome of the participatory design phase guided the development of the innovative BullStop mobile application in the fourth phase. The PD sessions were also used to identify essential features key to the app's acceptance amongst the target audience. The process of identifying these core features also highlighted the importance of reviewing the outputs of the UCD/PD iteratively; the co-designers amended the list of core features a number of times as their understanding of the proposed application increased. The mobile app provides users with a graphical interface to interact with a cloud backend that houses the ML models trained with the new dataset to predict cyberbullying and offensive language. As the system's visible component to the end-users, their perception of its usefulness and usability is critical for the application's success.

To investigate the system's performance in terms of its scalability and responsiveness, a range of experiments was conducted to evaluate these. The system demonstrated high responsiveness by returning prediction results through the Real-Time API in about a second and validated its ability to scale under heavy load by acquiring additional resources to cope with the increased load. Potential system obsolescence was shown to be successfully mitigated by the ability to use different classifiers dynamically and the generation of personalised classifiers by retraining the base classifier using ground truth supplied by users. A two-stage evaluation process was then conducted to evaluate the application's acceptability amongst end-users. This comprised a 'lab'-based study with a hand-selected number of users representing the stakeholders and a more extensive field-based study with beta users recruited via the Google Play app store. Findings revealed that the perceived ease of use and usefulness for the application are high, and its acceptability amongst the primary audience (i.e., young people) as well as other age groups (stakeholder groups) was similarly high. The app's acceptance by different age groups and stakeholders alike validates the use of UCD techniques for eliciting the requirements of the app as perceived by different types of stakeholders and the use of a PD approach to design the application.

A key difference between cyberbullying and other forms of online abuse is its repetitive nature. Cyberbullies are known to continuously torment their victims via social media. While the ML models trained and used in BullStop do not specifically address the repetitiveness of the detected instances of abuse, the application system that encompasses the ML model takes care of this by providing users with the option to automatically block repeat offenders and in so doing BullStop is able to tackle both repeated and isolated incidents of online abuse.

The three research questions posed in Chapter 1 as areas of inquiry for the research to explore are answered by the relevant scientific output from the work performed in the manner discussed below.

How can cyberbullying and online abuse be detected and prevented on social media platforms such that the key challenges of scalability, responsiveness, obsolescence and acceptability are adequately addressed?

The development of BullStop demonstrated the feasibility of performing scalable and responsive detection and prevention of cyberbullying on modern social media platforms. The evaluation experiments and studies conducted (see Chapter 7) demonstrated the system's responsiveness and ability to scale under heavy load. Furthermore, the extensible design that allows it to use different classifiers provides mitigation for system obsolescence. By creating a personalised classifier for users and allowing them to supplement the system's predictions with their own classifications, users provide the ML model with a means to gain insight into the way they communicate and use social media. This is a novel feature that has not been previously attempted in cyberbullying detection research and distinguishes this work from others.

The evaluation studies discovered that users found the app very useful and easy to use and in so doing demonstrated the app's acceptability among not only the primary target audience (i.e., young people) but also the other age groups that make up the secondary target audience.

What are stakeholders' needs and expectations for a cyberbullying prevention application and does the use of user-centred and participatory design approaches to design and develop the application results in an application that is an accurate

reflection of the stakeholders' expectations as measured by their perception of the tool's usability and usefulness?

The focus groups and interviews conducted (as reported in Chapter 4) provided the answers to the first part of this question in the form of the stakeholders' requirements for the app. The requirements solicited from both adult and young stakeholders during the knowledge elicitation activities were combined and further refined by the co-designers during the PD sessions, resulting in a prioritised list of essential features for the app (see Table 5.3 in Chapter 5).

With regards, the second part of this research question, the conceptualising and modelling of the stakeholders' requirements performed as part of the PD activities resulting in a design prototype that is a manifestation of the stakeholders' needs. This fed into the development of BullStop and resulted in an application that the human-based evaluation studies (see Chapter 7) scored highly on its perceived usability and usefulness across all users' age groups, with all surveyed users agreeing that the mobile application is well-designed for people of their age group. 88% of respondents in the field-based study provided a rating of very good or good for the app, and the remaining users (22%) rated the app as average. The UCD-led approach has, therefore, been very successful in ensuring that stakeholders' vision for the app has been accurately captured and translated via the PD activities into the developed mobile application.

What constitutes effective practice for engaging stakeholders in the user-centred research for the design and development of the cyberbullying prevention mobile application?

Reflections on the UCD/PD activities (Chapters 4 and 5) conducted as part of this research highlighted the importance of methodological flexibility when managing the engagement with the stakeholders and building rapport with participants from the outset to encourage active and productive participation. Recommendations based on these reflections were to:

- ensure adequate representation of all interested parties and stakeholders;
- build rapport with participants from the outset via frequent communication and providing them with access to information relevant to their participation;

- build flexibility into the design of the study protocols so that the study can progress with research integrity when faced with challenges;
- maximise opportunities to engage with participants while giving adequate consideration to the individual circumstances of the participants; and
- incorporate the means to empower participants within the study to encourage active participation

9.2 Contributions to Scientific Knowledge

This research identified and tackled (and thus contributed to scientific knowledge on) a number of challenges that negatively impact the availability of viable and effective cyberbullying detection and prevention tools, summarised as follows.

1. The survey of real-world cyberbullying prevention mobile apps conducted revealed the lack of practical applications for cyberbullying mitigation and prevention and highlighted the need for research programs such as this to address this issue. It provides a snapshot of the current real-world availability of mobile-based cyberbullying prevention applications and serves a resource to guide future researchers interested in developing similar tools.
2. The creation of a novel, large-scale multi-label dataset provides a robust resource for the training of ML models for the detection of online abuse and cyberbullying. The dataset was specifically designed to contain a large proportion of offensive content annotated for different forms of online abuse and cyberbullying. The dataset was successfully used to train a deep-learning model to detect online abuse in a corpus sourced from another online social network, and the high proportion of offensive content means it can be used without the need for oversampling techniques to boost the distribution of offensive content within the dataset. The dataset is publicly available for the use of other researchers in the field.
3. The machine learning experiments conducted to identify the best performing model and validate the generalisability of the created dataset demonstrated best practices in conducting performance evaluations for ML models and datasets and can aid future researchers embarking on similar endeavours.

4. Cyberbullying prevention systems are often developed in isolation without consultation with potential end-users. Subsequently, these tools struggle to gain acceptance amongst the intended audience due to not meeting users' expectations. This study is the first to engage all the identified stakeholder groups to devise automated strategies to mitigate and prevent cyberbullying and online abuse. This engagement made it possible to uncover valuable insight into what young people and adults feel about cyberbullying and how prevention and mitigation strategies can be automated in a cyberbullying prevention application to meet their expectations. Crucially, the study reinforced and extended existing scientific knowledge on cyberbullying and the effectiveness of existing prevention strategies and proposed new strategies to aid cyberbullying mitigation and prevention.
5. In recognising the importance of the developed system's impact on users and maintaining a focus on their interactions with the system throughout the development process, the application is assessed not only for its technical performance but also in terms of how its predictions can impact the lives of the users and their social networks. This study therefore established an approach for implementing ethics by design in the creation of AI-based systems. Future researchers can adopt a similar approach to ensure that the relevant ethical issues are duly considered when developing AI-based systems.
6. The design and implementation of Bullstop resulted in the creation of a novel cyberbullying detection and prevention application. BullStop uses a highly scalable and responsive cloud backend that can dynamically utilise different ML models and generate personalised cyberbullying detection classifiers for end-users. Its user interface is implemented as an Android application that was designed collaboratively with young people to ensure that it fully captures their needs and meets their expectations. BullStop is an impactful tool that provides real-world benefits, and the PD approach adopted for its creation offers other researchers a methodology for implementing research-backed systems that deliver real-world benefits. BullStop is freely available on the Google Play Store and has been installed over 400 times with an increasing active user base of just under one hundred.
7. The multi-dimensional evaluation study conducted to assess BullStop performance contributes to the knowledge on understanding an application's impact on its target

audience and its perceived usefulness and usability. The procedures and methods used in conducting the study can be adopted and adapted by future researchers to perform similar evaluation exercises.

8. Finally, reflections and recommendations from the research programme's use of UCD techniques to engage with stakeholders that ultimately resulted in the development of BullStop are provided as guidelines to assist future researchers in this domain.

9.3 Future Research

This research provided novel contributions to scientific knowledge and, in creating BullStop, produced a viable cyberbullying detection and prevention tool that has already impacted the lives of its users by helping them tackle online abuse and cyberbullying. During the course of the research, a number of areas have been identified as future research directions (themselves contributions to scientific knowledge that would not have been possible without the research) to extend the current work. These are discussed below.

1. **Enhancement of the Mobile Application.** Since the release of BullStop, the researcher has received emails (see FIGURE 9.1 for examples) directly from users asking for the inclusion of specific application features. When combined with the list of additional features and design refinements that were identified during the evaluation studies, the list of ways in which the app can be enhanced include:
 - supporting mobile devices running older versions of the Android operating system;
 - implementing the Message Checker feature as a virtual keyboard (similar to the ReThink and BBC Own It app discussed in Section 2.5);
 - integrating with other social media platforms like YouTube, Facebook and Instagram, even if the features available with these platforms are limited compared to the integration with Twitter;
 - providing an online chat feature by implementing a chatbot trained using advice content provided by bullying prevention organisations;

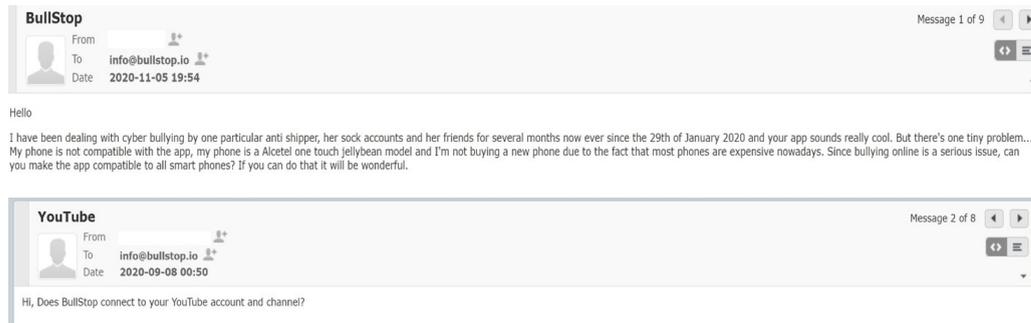


FIGURE 9.1 : Sample email messages from BullStop users

- reinstating the SMS/text feature that was disabled to achieve Google Play Store acceptance;
- providing access to more online resources on cyberbullying prevention and mitigation; and
- displaying daily motivational messages.

2. Explore the Use of Federated Learning by Leveraging a Network of Compacted Deep-Learning Models Running on the Mobile Client.

Federated Learning (Yang et al., 2019) is an area of machine learning that has recently been gaining attention due to its ability to utilise the local knowledge learned by multiple decentralised classifiers to improve the collective training of the entire system. Since the introduction of a federated learning architecture by Google in 2017 (McMahan and Ramage, 2017), it has only recently emerged as a viable technique for performing a range of machine learning tasks (Yang *et al.*, 2019; Li *et al.*, 2020). The research conducted can be extended to explore the possibility of accommodating the use of compacted ML models running on the mobile device instead of the cloud backend. Compact ML models like DistilBERT, ALBERT and MobileBERT (Sun *et al.*, 2020) are variants of deep-learning models such as RoBERTa and BERT that require a fraction of the computing resources of larger deep-learning models while retaining most of their performance. While DistilBERT was considered and discounted for use as a classifier in the BullStop system due to its poor performance compared to the other deep-learning models, a network of similar compact ML models in a federated learning configuration could potentially improve performance. Such a design would allow the mobile application to download a compacted deep-learning model from the cloud backend and train the model on the device using the ground truth provided by the user. The ground truth

data would stay local to the device and user, preserving the user's privacy, but the knowledge gained by the model would be shared across the network, thus improving the training of all the connected compact ML models.

3. Explore the use of computing cost as an evaluation metric for ML models .

The RoBERTa model used as the classifier in BullStop was selected based on its measured performance across several NLP experiments using a number of evaluation metrics and manual observation of its predictions on a sample of unseen tweets. While the deep learning models achieved better performance than the traditional ML models, they require significantly more computing resources and training time and for a system such as BullStop that is designed to avail online training to improve its performance, the time and the computing resources required to retrain models becomes a factor for consideration when assessing the system's long-term viability. A future research direction could be to develop a metric that assess the computing cost of a model as a factor of the GPU and training time.

4. Extension of the Dataset.

An obvious extension to the reported research would be to further increase the number of samples in the dataset. It would be especially beneficial if the numbers of minority classes like social exclusion, sarcasm and threat are improved to improve the performance of ML models trained using the dataset on the detection of these infrequent forms of online abuse. As the natural occurrence of these types of abuse on social media is low, samples from existing datasets like those of Oprea and Magdy (2019) and Rajadesingan *et al.* (2015) (for sarcasm-related tweets) could be extended by annotating them for the other labels used in the created dataset and the newly annotated documents added to the dataset created by this research program to improve the distribution of minority classes within the dataset.

5. Clustering Cyberbullying Attacks.

In the current implementation, the system's ML models identify online abuse and cyberbullying as isolated incidents. An extension of this could be associating each incident as part of a sequence of abuse that may involve multiple parties and determining the involved parties' roles. This could then be used to present a complete view of the online abuse and potentially used as evidence to aid legal prosecution.

6. Re-purposing BullStop for detecting and preventing other forms of online harms.

The same design that helps BullStop mitigate obsolescence can be easily re-purposed to expand its use into other areas of online harms such as sexual grooming, stalking and 'catfishing' (using a fake online identity to start romantic relations). Different ML models can be created by training on data sourced and annotated for different types of online harms and in this way, the system can be used to tackle these other forms of online dangers. Additionally, the UI can be amended such that instead of a fixed set of labels as is currently the case with online abuse detection and prevention, predicted labels are dynamically created based on the ML models' outputs and associated with the relevant messages.

References

- Marieke Vanden Abeele and Rozane De Cock. 2013. Cyberbullying by mobile phone among adolescents: The role of gender and peer group status. *Communications*, 38:107–118.
- Çiğdem Aci, Eren ÇÜRÜK, and Esra Saraç EŞSİZ. 2019. Automatic detection of cyberbullying in formspring.me, myspace and youtube social networks. *Turkish Journal of Engineering*, 3:168–178.
- Patricia Agatston, Robin Kowalski, and Susan Limber. 2012. Youth views on cyberbullying. *Cyberbullying prevention and response: Expert perspectives*, pages 57–71.
- Sweta Agrawal and Amit Awekar. 2018. Deep learning for detecting cyberbullying across multiple social media platforms. volume 10772 LNCS, pages 141–153. Springer Verlag.
- Mohammed Ali Al-Garadi, Kasturi Dewi Varathan, and Sri Devi Ravana. 2016. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network. *Computers in Human Behavior*, 63:433–443.
- Lizette Alvarez. 2013. Girl's suicide points to rise in apps used by cyberbullies.
- Silvio Amir, Byron C. Wallace, Hao Lyu, and Paula Carvalho Mário J. Silva. 2016. Modelling context with user embeddings for sarcasm detection in social media. *CoNLL 2016 - 20th SIGNLL Conference on Computational Natural Language Learning, Proceedings*, pages 167–177.
- Mukul Anand and R. Eswari. 2019. Classification of abusive comments in social media using deep learning. pages 974–977. Institute of Electrical and Electronics Engineers Inc.

- Laksmi Anindyati, Ayu Purwarianti, and Ade Nursanti. 2019. Optimizing deep learning for detection cyberbullying text in Indonesian language. Institute of Electrical and Electronics Engineers Inc.
- Zahra Ashktorab and Jessica Vitak. 2016. Designing cyberbullying mitigation and prevention solutions through participatory design with teenagers. pages 3895–3905. Association for Computing Machinery.
- Jason Aten. 2020. Apple's standoff with developers over the app store puts it on the wrong side of innovation.
- Dzmitry Bahdanau, Tom Bosc, Stanisaw Jastrzbski, Edward Grefenstette, Pascal Vincent, and Yoshua Bengio. 2017. Learning to compute word embeddings on the fly.
- Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling.
- Anna C. Baldry, David P. Farrington, and Anna Sorrentino. 2016. Cyberbullying in youth: A pattern of disruptive behaviour. *Psicologia Educativa*, 22:19–26.
- Vijay Banerjee, Jui Telavane, Pooja Gaikwad, and Pallavi Vartak. 2019. Detection of cyberbullying using deep neural network. pages 604–607. Institute of Electrical and Electronics Engineers Inc.
- Rosaline Barbour and Jenny Kitzinger. 1998. *Developing focus group research: politics, theory and practice*. Sage.
- Christopher P. Barlett. 2015. Anonymously hurting others online: The effect of anonymity on cyberbullying frequency. *Psychology of Popular Media Culture*, 4:70–79.
- Sheri Bauman. 2010. Cyberbullying in a rural intermediate school: An exploratory study. *The Journal of Early Adolescence*, 30:803–833.
- Sheri Bauman. 2015. Types of cyberbullying.
- BBC News. 2013. Hannah Smith death: Father says daughter was victim of cyberbullies.
- BBC News. 2017. Hunt challenges social media giants on cyber-bullying.
- BBC News. 2018. Social media firms 'failing' to tackle cyber-bullying.
- BBC News. 2019. Snapchat bullying: Adults 'can't keep up with technology'.

- BBC News. 2020. 7 stars who have personal experiences of online bullying.
- Linda Beckman, Curt Hagquist, and Lisa Hellström. 2013. Discrepant gender patterns for cyberbullying and traditional bullying – an analysis of swedish adolescent data. *Computers in Human Behavior*, 29:1896–1903.
- Emilio Bellini, Gerardo Canfora, Félix García, Mario Piattini, and Corrado Aaron Visaggio. 2005. Pair designing as practice for enforcing and diffusing design knowledge. *Journal of Software Maintenance and Evolution: Research and Practice*, 17:401–423.
- Laura Benton, Hilary Johnson, Emma Ashwin, Mark Brosnan, and Beate Grawemeyer. 2012. Developing ideas: Supporting children with autism within a participatory design team. pages 2599–2608.
- Sofia Berne, Ann Frisé, and Johanna Kling. 2014. Appearance-related cyberbullying: A qualitative investigation of characteristics, content, reasons, and effects. *Body Image*, 11:527–533.
- Lesli Biediger-Friedman, Monica Silva, and Kenneth Smith. 2018. A focus group study observing maternal intention to use a wic education app. *American Journal of Health Behavior*, 42:110–123.
- Laura P. v Bosque and Sara Elena Garza. 2014. Aggressive text detection for cyberbullying. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8856:221–232.
- Leanne Bowler, Eleanor Mattern, and Cory Knobel. 2014. Developing design interventions for cyberbullying: A narrative-based participatory approach. iSchools.
- Jordan L Boyd-Graber, Sonya S Nikolova, Karyn A Moffatt, Kenrick C Kin, Joshua Y Lee, Lester W Mackey, Marilyn M Tremaine, and Maria M Klawe. 2006. Participatory design with proxies: developing a desktop-pda system to support people with aphasia. pages 151–160.
- Julia Brailovskaia, Tobias Teismann, and Jürgen Margraf. 2018. Cyberbullying, positive mental health and suicide ideation/behavior. *Psychiatry Research*, 267:240–242.
- I. Bratko. 1997. Machine learning: Between accuracy and interpretability.

- Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. 2017. *Classification and regression trees*. CRC Press.
- Uwe Bretschneider, Thomas Wöhner, and Ralf Peters. 2014. Detecting online harassment in social networks. page 1.
- David Van Bruwaene, Qianjia Huang, and Diana Inkpen. 2020. A multi-platform dataset for detecting cyberbullying in social media. *Language Resources and Evaluation*, pages 1–24.
- Seok Jun Bu and Sung Bae Cho. 2018. A hybrid deep learning system of cnn and lrcn to detect cyberbullying from sns comments. volume 10870 LNAI, pages 561–572. Springer Verlag.
- Ana Paula Caetano, Isabel Freire, Ana Margarida Veiga Simão, Maria José D. Martins, and Maria Teresa Pessoa. 2016. Emotions in cyberbullying: A study with portuguese teenagers. *Educacao e Pesquisa*, 42:199–212.
- Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. pages 2506–2515.
- Fethi Calisir and Ferah Calisir. 2004. The relation of interface usability characteristics, perceived usefulness, and perceived ease of use to end-user satisfaction with enterprise resource planning (erp) systems. *Computers in Human Behavior*, 20:505–515.
- Esther Calvete, Izaskun Orue, Ana Estévez, Lourdes Villardón, and Patricia Padilla. 2010. Cyberbullying in adolescents: Modalities and aggressors' profile. *Computers in Human Behavior*, 26:1128–1135.
- Antonio Calvo-Morata, Dan Cristian Rotaru, Cristina Alonso-Fernandez, Manuel Freire, Ivan Martinez-Ortiz, and Baltasar Fernandez-Manjon. 2018. Validation of a cyberbullying serious game using game analytics. *IEEE Transactions on Learning Technologies*.
- Sonia Camacho, Khaled Hassanein, and Milena Head. 2018. Cyberbullying impacts on victims' satisfaction with information and communication technologies: The role of perceived cyberbullying severity. *Information and Management*, 55:494–507.

- Xiongfei Cao, Ali Nawaz Khan, Ahsan Ali, and Naseer Abbas Khan. 2019. Consequences of cyberbullying and social overload while using snss: A study of users' discontinuous usage behavior in snss. *Information Systems Frontiers*.
- Wanda Cassidy, Chantal Faucher, and Margaret Jackson. 2018. What parents can do to prevent cyberbullying: Students' and educators' perspectives. *Social Sciences*, 7:251.
- Edgar Cebolledo and Olga De Troyer. 2015. Modelling social network interactions in games. pages 82–88.
- Nidhi Chandra, Sunil Kumar Khatri, and Subhranil Som. 2018. Cyberbullying detection using recursive neural network through offline repository. pages 748–754. Institute of Electrical and Electronics Engineers Inc.
- Ross Chapman. 2018. Pair design: Guidelines for trying it out in your next design sprint.
- Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. pages 13–22. Association for Computing Machinery, Inc.
- Vikas S. Chavan and S. S. Shylaja. 2015. Machine learning approach for detection of cyber-aggressive comments by peers on social media network. pages 2354–2358. Institute of Electrical and Electronics Engineers Inc.
- Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.
- Hsin-Yu Chen and Cheng-Te Li. 2020. Henin: Learning heterogeneous neural interaction networks for explainable cyberbullying detection on social media.
- Junyi Chen, Shankai Yan, and Ka Chun Wong. 2018. Verbal aggression detection on twitter comments: convolutional neural network for short-text sentiment analysis. *Neural Computing and Applications*, pages 1–10.
- Lu Cheng, Jundong Li, Yasin Silva, Deborah Hall, and Huan Liu. 2019. Pi-bully: personalized cyberbullying detection with peer influence. pages 5829–5835. AAAI Press.

- Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation.
- Kamal Choudhary, Marnik Berx, Jie Jiang, Ruth Pachter, Dirk Lamoen, and Francesca Tavazza. 2019. *Accelerated Discovery of Efficient Solar-cell Materials using Quantum and Machine-learning Methods*.
- Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. volume 7.
- Katrien Van Cleemput, Heidi Vandebosch, and Sara Pabian. 2014. Personal characteristics and contextual factors that determine "helping," "joining in," and "doing nothing" when witnessing cyberbullying. *Aggressive Behavior*, 40:383–396.
- Jane Clemensen, Simon B. Larsen, Morten Kyng, and Marit Kirkevold. 2007. Participatory design in health sciences: Using cooperative experimental methods in developing health services and computer technology. *Qualitative Health Research*, 17:122–130.
- Vítor Alexandre Coelho, Vanda Sousa, Marta Marchante, Patrícia Brás, and Ana Maria Romão. 2016. Bullying and cyberbullying in portugal: Validation of a questionnaire and analysis of prevalence. *School Psychology International*, 37:223–239.
- David Cohen. 2015. Facebook changes definition of monthly active users.
- Wendy M Craig and Yossi Harel. 2001. Bullying, physical fighting and victimization. *Young people's health in context: International report from the HBSC*, 2:133–144.
- Sławomir Dadas, Michał Perelkiewicz, and Rafał Poświata Poświata. 2020. Pre-training polish transformer-based language models at scale a preprint.
- Maral Dadvar and Kai Eckert. 2020. Cyberbullying detection in social networks using deep learning based models. volume 12393 LNCS, pages 245–255. Springer Science and Business Media Deutschland GmbH.
- Maral Dadvar, F M G de Jong, Roeland Ordelman, and Dolf Trieschnigg. 2012a. Improved cyberbullying detection using gender information. University of Ghent.

- Maral Dadvar and Franciska De Jong. 2012. Cyberbullying detection: A step toward a safer internet yard. pages 121–125.
- Maral Dadvar, Roeland Ordelman, Franciska De Jong, and Dolf Trieschnigg. 2012b. Towards user modelling in the combat against cyberbullying. volume 7337 LNCS, pages 277–283.
- Maral Dadvar, Dolf Trieschnigg, and Franciska De Jong. 2013. Expert knowledge for automatic detection of bullies in social networks. pages 57–64.
- Thomas Davidson, Dana Warmesley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language.
- Alexandre Davis, Adriano Veloso, Altigran Soares, Alberto Laender, and Wagner Meira Jr. 2012. Named entity disambiguation in streaming data. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 815–824, Jeju Island, Korea. Association for Computational Linguistics.
- Francine Dehue. 2013. Cyberbullying research: New perspectives and alternative methodologies. introduction to the special issue. *Journal of Community Applied Social Psychology*, 23:1–6.
- Francine Dehue, Catherine Bolman, and Trijntje Völlink. 2008. Cyberbullying: Youngsters' experiences and parental perception. *Cyberpsychology and Behavior*, 11:217–223.
- A. DeSmet, S. Bastiaensens, K. Van Cleemput, K. Poels, H. Vandebosch, G. Deboutte, L. Herrewijn, S. Malliet, S. Pabian, F. Van Broeckhoven, O. De Troyer, G. Deglorie, S. Van Hoecke, K. Samyn, and I. De Bourdeaudhuij. 2018a. The efficacy of the friendly attac serious digital game to promote prosocial bystander behavior in cyberbullying among young adolescents: A cluster-randomized controlled trial. *Computers in Human Behavior*, 78:336–347.
- A. DeSmet, M. Rodelli, M. Walrave, B. Soenens, G. Cardon, and I. De Bourdeaudhuij. 2018b. Cyberbullying and traditional bullying involvement among heterosexual and non-heterosexual adolescents, and their associations with age and gender. *Computers in Human Behavior*, 83:254–261.
- Ann DeSmet, Sara Bastiaensens, Katrien Van Cleemput, Karolien Poels, Heidi Vandebosch, Greet Cardon, and Ilse De Bourdeaudhuij. 2016. Deciding whether to

- look after them, to like it, or leave it: A multidimensional analysis of predictors of positive and negative bystander behavior in cyberbullying among adolescents. *Computers in Human Behavior*, 57:398–415.
- Sebastien Destercke. 2014. Multilabel prediction with probability sets: The hamming loss case. volume 443 CCIS, pages 496–505. Springer Verlag.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.
- Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems*, 2.
- Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying.
- Ditch The Label. 2020. What is cyberbullying?
- Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. volume 7, pages 67–73. Association for Computing Machinery, Inc.
- Julian J. Dooley, Jacek Pyzalski, and Donna Cross. 2009. Cyberbullying versus face-to-face bullying: A theoretical and conceptual review.
- Jennifer Doty, Amy Gower, Renee Sieving, Shari Plowman, and Barbara McMorris. 2018. Cyberbullying victimization and perpetration, connectedness, and monitoring of online activities: Protection from parental figures. *Social Sciences*, 7:265.
- Stephan Dreiseitl and Lucila Ohno-Machado. 2002. Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, 35:352–359.
- Sigal Eden, Tali Heiman, and Dorit Olenik-Shemesh. 2013. Teachers' perceptions, beliefs and concerns about cyberbullying. *British Journal of Educational Technology*, 44:1036–1052.
- Lynne Edwards, April Edwards Kontostathis, and Christina Fisher. 2016. Cyberbullying, race/ethnicity and mental health outcomes: A review of the literature.

- Paz Elipe, María de la Oliva Muñoz, and Rosario Del Rey. 2018. Homophobic bullying and cyberbullying: Study of a silenced problem. *Journal of Homosexuality*, 65:672–686.
- Abdelrahim Elmadany, Chiyu Zhang, Muhammad Abdul-Mageed, and Azadeh Hashemi. 2020. Leveraging affective bidirectional transformers for offensive language detection. *ArXiv*.
- Jake Evans. 2018. Fake instagram accounts being used by kids to 'destroy reputations', esafety commissioner says - abc news.
- Facebook. 2018. An update on our plans to restrict data access on facebook.
- Gunter Fahrnberger, Deveeshree Nayak, Venkata Swamy Martha, and Srinu Ramaswamy. 2014. Safechat: A tool to shield children's communication from explicit messages. pages 80–86. IEEE Computer Society.
- Amanda E. Fahy, Stephen A. Stansfeld, Melanie Smuk, Neil R. Smith, Steven Cummins, and Charlotte Clark. 2016. Longitudinal associations between cyberbullying involvement and adolescent mental health. *Journal of Adolescent Health*, 59:502–509.
- Mingyue Fan, Liyue Yu, and Leanne Bowler. 2016. Feelbook: A social media app for teens designed to foster positive online behavior and prevent cyberbullying. volume 07-12-May-2016, pages 1187–1192. Association for Computing Machinery.
- Kostas A. Fanti, Andreas G. Demetriou, and Veronica V. Hawa. 2012. A longitudinal study of cyberbullying: Examining risk and protective factors. *European Journal of Developmental Psychology*, 9:168–181.
- Christiane Floyd. 1993. Steps - a methodical approach to pd. *Communications of the ACM*, 36:83–84.
- Mairéad Foody, Lian McGuire, Seffetullah Kuldass, and James O'Higgins Norman. 2019. Friendship quality and gender differences in association with cyberbullying involvement and psychological well-being. *Frontiers in Psychology*, 10.
- Yee Jang Foong and Mourad Oussalah. 2017. Cyberbullying system detection and analysis. volume 2017-January, pages 40–46. Institute of Electrical and Electronics Engineers Inc.

- Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior.
- Christopher Frauenberger, Judith Good, and Wendy Keay-Bright. 2011. Designing technology for children with special needs: bridging perspectives through participatory design. *CoDesign*, 7:1–28.
- Minghui Gao, Xu Zhao, and Mark McJunkin. 2016. Adolescents' experiences of cyberbullying: Gender, age and reasons for not reporting to adults. *International Journal of Cyber Behavior, Psychology and Learning (IJCPL)*, 6:13–27.
- Maite Garaigordobil and Vanesa Martínez-Valderrey. 2018. Technological resources to prevent cyberbullying during adolescence: The cyberprogram 2.0 program and the cooperative cybereduca 2.0 videogame. *Frontiers in Psychology*, 9.
- Amy W. Gatian. 1994. Is user satisfaction a valid measure of system effectiveness? *Information and Management*, 26:119–131.
- Anita Gibbs. 1997. Focus groups. *Social research update*, 19:1–8.
- Athanasia Gkiomisi, Maria Gkrizioti, Athina Gkiomisi, Dimitrios A. Anastasilakis, and Panagiotis Kardaras. 2017. Cyberbullying among greek high school adolescents. *Indian Journal of Pediatrics*, 84:364–368.
- R. Matthew. Gladden, Alana M. Vivolo-Kantor, Merle E. Hamburger, and Corey D. Lumpkin. 2014. Bullying surveillance among youths : uniform definitions for public health and recommended data elements, version 1.0.
- Deborah Goebert, Iwalani Else, Courtenay Matsu, Jane Chung-Do, and Janice Y. Chang. 2011. The impact of cyberbullying on substance use and mental health in a multiethnic sample. *Maternal and Child Health Journal*, 15:1282–1286.
- Google. 2020. Use of sms or call log permission groups.
- Jon D. Goss and Thomas R. Leinbach. 1996. Focus groups as alternative research practice: Experience with transmigrants in indonesia. *Area*, 28:115–123.

- Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18:602–610.
- Judith Gregory. 2003. Scandinavian approaches to participatory design. *International Journal of Engineering Education*, 19:62–74.
- Lucy Griesel, Linda R. Finger, Gawaiian H. Bodkin-Andrews, Rhonda G. Craven, and Alexander Seeshing Yeung. 2012. Uncovering the structure of and gender and developmental differences in cyber bullying.
- Patrick J Grother, Mei L Ngan, and Kayee K Hanaoka. 2019. Face recognition vendor test (frvt) part 3: Demographic effects.
- Suzanne Guerin and Eilis Hennessy. 2002. Pupils' definitions of bullying. *European Journal of Psychology of Education*, 17:249–261.
- Aabhaas Gupta, Wenxi Yang, Divya Sivakumar, Yasin Silva, Deborah Hall, and Maria Nardini Barioni. 2020. Temporal properties of cyberbullying on instagram. pages 576–583. Association for Computing Machinery.
- M Guzman-Silverio, A Balderas-Paredes, and A P López-Monroy. 2020. Transformers and data augmentation for aggressiveness detection in mexican spanish.
- Manuel Gámez-Guadix, Erika Borrajo, and Carmen Almendros. 2016. Risky online behaviors among adolescents: Longitudinal relations among problematic internet use, cyberbullying perpetration, and meeting strangers online. *Journal of Behavioral Addictions*, 5:100–107.
- Manuel Gámez-Guadix, Izaskun Orue, Peter K. Smith, and Esther Calvete. 2013. Longitudinal and reciprocal relations of cyberbullying with depression, substance use, and problematic internet use among adolescents. *Journal of Adolescent Health*, 53:446–452.
- Uğur Gündüz. 2017. The effect of social media on identity construction. *Mediterranean Journal of Social Sciences*, 8:85.
- Lilit Hakobyan, Jo Lumsden, and Dympna O'Sullivan. 2014. Older adults with amd as co-designers of an assistive mobile application. *International Journal of Mobile Human Computer Interaction (IJMHCI)*, 6:54–70.

- Elizabeth J. Halcomb, Leila Gholizadeh, Michelle DiGiacomo, Jane Phillips, and Patricia M. Davidson. 2007. Literature review: Considerations in undertaking focus group research with culturally and linguistically diverse groups.
- Ong Chee Hang and Halina Mohamed Dahlan. 2019. Cyberbullying lexicon for social media. volume December-2019. IEEE Computer Society.
- J. A. Hanley and B. J. McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143:29–36.
- Helena Harder, Patrick Holroyd, Lynn Burkinshaw, Phil Watten, Charles Zammit, Peter R Harris, Anna Good, and Val Jenkins. 2017. A user-centred approach to developing bwell, a mobile app for arm and shoulder exercises after breast cancer treatment.
- Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. 2008. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. pages 1322–1328. IEEE.
- Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2018. Automatic detection of cyberbullying in social media text. *PLOS ONE*, 13:e0203794.
- Kathleen Hemenway. 1982. Psychological issues in the use of icons in command menus. pages 20–25. Association for Computing Machinery.
- Starr Roxanne Hiltz and Kenneth Johnson. 1990. User satisfaction with computer-mediated communication systems. *Management Science*, 36:739–764.
- Sameer Hinduja and Justin W. Patchin. 2008. Cyberbullying: An exploratory analysis of factors related to offending and victimization. *Deviant Behavior*, 29:129–156.
- Sameer Hinduja and Justin W. Patchin. 2010. Bullying, cyberbullying, and suicide. *Archives of Suicide Research*, 14:206–221.
- Sameer Hinduja and Justin W. Patchin. 2012. Cyberbullying: Neither an epidemic nor a rarity. *European Journal of Developmental Psychology*, 9:539–543.
- Sameer Hinduja and Justin W. Patchin. 2013. Social influences on cyberbullying behaviors among middle and high school students. *Journal of Youth and Adolescence*, 42:711–722.

- Sameer Hinduja and Justin W Patchin. 2019. Connecting adolescent suicide to the severity of bullying and cyberbullying. *Journal of School Violence*, 18:333–346.
- Tin Kam Ho. 1995. Random decision forests. volume 1, pages 278–282. IEEE Computer Society.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9:1735–1780.
- Dianne L. Hoff and Sidney N. Mitchell. 2009. Cyberbullying: Causes, effects, and remedies. *Journal of Educational Administration*, 47:652–665.
- Jun Sung Hong, Jungup Lee, Dorothy L. Espelage, Simon C. Hunter, Desmond Upton Patton, and Tyrone Rivers. 2016. Understanding the correlates of face-to-face and cyberbullying victimization among u.s. adolescents: A social-ecological analysis. *Violence and Victims*, 31:638–663.
- Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Poster: Detection of cyberbullying in a mobile social network: Systems issues. page 481. Association for Computing Machinery, Inc.
- Homa Hosseinmardi, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2016. Prediction of cyberbullying incidents in a media-based social network. pages 186–192. Institute of Electrical and Electronics Engineers Inc.
- Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. pages 591–598.
- Yulin Hswen, Lauren Rubenzahl, and David S. Bickham. 2014. Feasibility of an online and mobile videogame curriculum for teaching children safe and healthy cellphone and internet behaviors. *Games for Health Journal*, 3:252–259.
- Qianjia Huang, Vivek Kumar Singh, and Pradeep Kumar Atrey. 2014. Cyber bullying detection using social and textual analysis. pages 3–6. ACM.
- Sofia Hussain. 2010. Empowering marginalised children in developing countries through participatory design processes. *CoDesign*, 6:99–117.
- Mai Ibrahim, Marwan Torki, and Nagwa El-Makky. 2019. Imbalanced toxic comments classification using data augmentation and deep learning. pages 875–878. Institute of Electrical and Electronics Engineers Inc.

- IDC. 2020. Smartphone market share.
- Begoña Iranzo, Sofía Buelga, María Jesús Cava, and Jessica Ortega-Barón. 2019. Cyberbullying, psychosocial adjustment, and suicidal ideation in adolescence. *Psychosocial Intervention*, 28:75–81.
- Ole Sejer Iversen, Rachel Charlotte Smith, and Christian Dindler. 2017. Child as protagonist: Expanding the role of children in participatory design re-humanising automated-decision making view project. *dl.acm.org*, pages 27–37.
- Amit Kumar Jena, Aman Sinha, and Rohit Agarwal. 2020. C-net: Contextual network for sarcasm detection. pages 61–66.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification.
- Thomas Jäger, João Amado, Armanda Matos, and Teresa Pessoa. 2010. Analysis of experts' and trainers' views on cyberbullying. *Australian Journal of Guidance and Counselling*, 20:169–181.
- Kaggle. 2012. Detecting insults in social commentary.
- Kaggle. 2018. Toxic comment classification challenge.
- Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. volume 1, pages 655–665. Association for Computational Linguistics (ACL).
- Saloni Mahesh Kargutkar and Vidya Chitre. 2020. A study of cyberbullying detection using machine learning techniques. pages 734–739. Institute of Electrical and Electronics Engineers Inc.
- Finn Kensing and Jeanette Blomberg. 1998. Participatory design: Issues and concerns. *Computer Supported Cooperative Work*, 7:167–185.
- Aye Thazin Khine, Yu Mon Saw, Zaw Ye Htut, Cho Thet Khaing, Htin Zaw Soe, Kyu Kyu Swe, Thinzar Thike, Hein Htet, Thu Nandar Saw, and Su Myat Cho. 2020. Assessing risk factors and impact of cyberbullying victimization among university students in myanmar: A cross-sectional study. *PloS one*, 15:e0227051.

- Naresh Khuriwal and Nidhi Mishra. 2018. Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm. pages 1–5. Institute of Electrical and Electronics Engineers Inc.
- Soyeon Kim, Michael H. Boyle, and Katholiki Georgiades. 2017. Cyberbullying victimization and its association with health across the life course: A canadian population study. *Canadian Journal of Public Health*, 108:e468–e474.
- Soyeon Kim, Scott R. Colwell, Anna Kata, Michael H. Boyle, and Katholiki Georgiades. 2018. Cyberbullying victimization and adolescent mental health: Evidence of differential effects by sex and mental health problem type. *Journal of Youth and Adolescence*, 47:661–672.
- Jenny Kitzinger. 1995. Qualitative research: introducing focus groups. *Bmj*, 311:299–302.
- April Kontostathis, Kelly Reynolds, Andy Garron, and Lynne Edwards. 2013. Detecting cyberbullying: Query terms and techniques. volume volume, pages 195–204. Association for Computing Machinery.
- Robin M. Kowalski and Susan P. Limber. 2007. Electronic bullying among middle school students. *Journal of Adolescent Health*, 41.
- Klaus Krippendorff. 2011. Computing krippendorff’s alpha-reliability. *Departmental Papers (ASC)*.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84–90.
- Richard A Krueger. 2014. *Focus groups: A practical guide for applied research*. Sage publications.
- Shamanth Kumar, Reza Zafarani, and Huan Liu. 2011. Understanding user migration patterns in social media.
- Kirti Kumari, Jyoti Prakash Singh, Yogesh K. Dwivedi, and Nripendra P. Rana. 2019. Aggressive social media post detection system containing symbolic images. volume 11701 LNCS, pages 415–424. Springer Verlag.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv*.

- James A. Landay and Brad A. Myers. 2001. Sketching interfaces: toward more human interface design. *Computer*, 34:56–64.
- Danielle M. Law, Jennifer D. Shapka, José F. Domene, and Monique H. Gagné. 2012. Are cyberbullies really bullies? an investigation of reactive and proactive online aggression. *Computers in Human Behavior*, 28:664–672.
- Changho Lee and Namin Shin. 2017. Prevalence of cyberbullying and predictors of cyberbullying perpetration among korean adolescents. *Computers in Human Behavior*, 68:352–358.
- Ho Suk Lee, Hong Rae Lee, Jun U. Park, and Yo Sub Han. 2018. An abusive text detection system based on enhanced abusive and non-abusive word lists. *Decision Support Systems*, 113:22–31.
- Sebastian Lehrig, Hendrik Eikerling, and Steffen Becker. 2015. Scalability, elasticity, and efficiency in cloud computing: A systematic literature review of definitions and metrics. pages 83–92. Association for Computing Machinery, Inc.
- Pawel Lempa, Michal Ptaszynski, and Fumito Masui. 2015. Cyberbullying blocker application for android.
- Qing Li. 2007. Bullying in the new playground: Research into cyberbullying and cyber victimisation cyberbullying and cyber victimisation.
- Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2020. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37:50–60.
- Wanqi Li. 2019. A design approach for automated prevention of cyberbullying using language features on social media. pages 87–91. Institute of Electrical and Electronics Engineers Inc.
- Helen Lianos and Andrew McGrath. 2018. Can the general theory of crime and general strain theory explain cyberbullying perpetration? *Crime Delinquency*, 64:674–700.
- Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5:1–184.

- Huan Liu and Rudy Setiono. 1995. Chi2: feature selection and discretization of numeric attributes. pages 388–391. IEEE.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.
- Cherlynn Low. 2018. Hey google, android actually does stifle competition.
- Jo Lumsden, Lilit Hakobyan, Rock Leung, and Dympna O’Sullivan. 2017. Disabilities: Assistive technology design.
- Minh Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. pages 1412–1421. Association for Computational Linguistics (ACL).
- Antonio López-Martínez, José Antonio García-Díaz, Rafael Valencia-García, and Antonio Ruiz-Martínez. 2019. Cyberdect. a novel approach for cyberbullying detection on twitter. volume 1124 CCIS, pages 109–121. Springer.
- Peter J. R. Macaulay, Lucy R. Betts, James Stiller, and Blerina Kellezi. 2020. “it’s so fluid, it’s developing all the time”: pre-service teachers’ perceptions and understanding of cyberbullying in the school environment. *Educational Studies*, 46:590–606.
- Colin MacDougall and Elizabeth Fudge. 2001. Planning and recruiting the sample for focus groups and in-depth interviews.
- Hana Machackova and Jan Pfetsch. 2016. Bystanders’ responses to offline bullying and cyberbullying: The role of empathy and normative beliefs about aggression. *Scandinavian Journal of Psychology*, 57:169–176.
- Juan Manuel Machimbarrena and Maite Garaigordobil. 2018. Prevalence of bullying and cyberbullying in the last stage of primary education in the basque country. *Spanish Journal of Psychology*, 21.
- Martin Maguire and Nigel Bevan. 2002. User requirements analysis. pages 133–148.
- E. Makri-Botsari and G. Karagianni. 2014. Cyberbullying in greek adolescents: The role of parents. *Procedia - Social and Behavioral Sciences*, 116:3241–3253.

- Aditya Malte and Pratik Ratadiya. 2019. Multilingual cyber abuse detection using advanced transformer architecture. volume 2019-October, pages 784–789. Institute of Electrical and Electronics Engineers Inc.
- Juan F. Mancilla-Caceres, Dorothy Espelage, and Eyal Amir. 2015. A computer game-based method for studying bullying and cyberbullying. *Journal of School Violence*, 14:66–86.
- Juan F. Mancilla-Caceres, Wen Pu, Eyal Amir, and Dorothy Espelage. 2012. A computer-in-the-loop approach for detecting bullies in the classroom. volume 7227 LNCS, pages 139–146.
- Amrita Mangaonkar, Allenous Hayrapetian, and Rajeev Raje. 2015. Collaborative detection of cyberbullying behavior in twitter data. volume 2015-June, pages 611–616. IEEE Computer Society.
- Gary Marcus. 2018. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.
- Catherine Marshall and Gretchen B Rossman. 2014. *Designing qualitative research*. Sage publications.
- María Carmen Martínez-Monteagudo, Beatriz Delgado, Ángela Díaz-Herrero, and José Manuel García-Fernández. 2020. Relationship between suicidal thinking, anxiety, depression and stress in university students who are victims of cyberbullying. *Psychiatry Research*, 286:112856.
- Michael Massimi and Ronald Baecker. 2006. Participatory design process with older users.
- Ali Al Mazari. 2013. Cyber-bullying taxonomies: Definition, forms, consequences and mitigation strategies. pages 126–133. IEEE.
- Isabella McLafferty. 2004. Focus group interviews as a data collecting strategy.
- Brendan McMahan and Daniel Ramage. 2017. Google ai blog: Federated learning: Collaborative machine learning without centralized training data.
- Brenna McNally, Priya Kumar, Chelsea Hordatt, Matthew Louis Mauriello, Shalmali Naik, Leyla Norooz, Alazandra Shorter, Evan Golub, and Allison Druin. 2018. Co-designing mobile online safety applications with children. volume 2018-April, pages 1–9. Association for Computing Machinery.

- Charles E Metz. 1978. Basic principles of roc analysis. volume 8, pages 283–298. WB Saunders.
- Josephina Mikka-Muntuumo, Anicia Peters, and Hussin Jazri. 2018. Cyberbullet – share your story: An interactive game for stimulating awareness on the harm and negative effects of the internet. pages 287–290. Association for Computing Machinery.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. International Conference on Learning Representations, ICLR.
- Claire P Monks, Jess Mahdavi, and Katie Rix. 2016. The emergence of cyberbullying in childhood: Parent and teacher perspectives. *Psicología Educativa*, 22:39–48.
- Neema Moraveji, Jason Li, Jiarong Ding, Patrick O'Kelley, and Suze Woolf. 2007. Comicboarding: Using comics as proxies for participatory design with children. pages 1371–1374. Association for Computing Machinery.
- Megan A. Moreno, Nina Suthamjariya, and Ellen Selkie. 2018. Stakeholder perceptions of cyberbullying cases: Application of the uniform definition of bullying. *Journal of Adolescent Health*, 62:444–449.
- Francesca Moretti, Liesbeth van Vliet, Jozien Bensing, Giuseppe Deledda, Mariangela Mazzi, Michela Rimondini, Christa Zimmermann, and Ian Fletcher. 2011. A standardized approach to qualitative content analysis of focus group discussions from different countries. *Patient Education and Counseling*, 82:420–428.
- Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. A bert-based transfer learning approach for hate speech detection in online social media. volume 881 SCI, pages 928–940. Springer.
- Michael J Muller and Sarah Kuhn. 1993. Participatory design. *Communications of the ACM*, 36:24–28.
- Myriam Munezero, Calkin Suero Montero, Tuomo Kakkonen, Erkki Sutinen, Maxim Mozgovoy, and Vitaly Klyuev. 2014. Automatic detection of antisocial behaviour in texts. *Informatica*, 38.

- Myriam Munezero, Maxim Mozgovoy, Tuomo Kakkonen, Vitaly Klyuev, and Erkki Sutinen. 2013. Antisocial behavior corpus for harmful language detection. pages 261–265. IEEE.
- Vinita Nahar, Sanad Al-Maskari, Xue Li, and Chaoyi Pang. 2014. Semi-supervised learning for cyberbullying detection in social networks. volume 8506 LNCS, pages 160–171. Springer Verlag.
- Vinita Nahar, Xue Li, and Chaoyi Pang. 2013. An effective approach for cyberbullying detection. *Communications in Information Science and Management Engineering*, 3:238.
- G NaliniPriya and M Asswini. 2015. A dynamic cognitive system for automatic detection and prevention of cyber-bullying attacks. *ARPN Journal of Engineering and Applied Science*, 10.
- B. Sri Nandhini and J. I. Sheeba. 2015. Online social network bullying detection using intelligence techniques. volume 45, pages 485–492. Elsevier B.V.
- Jordana N. Navarro and Jana L. Jasinski. 2013. Why girls? using routine activities theory to predict cyberbullying experiences between girls and boys. *Women and Criminal Justice*, 23:286–303.
- Jakob Nielsen. 1994. Heuristic evaluation, w: Nielsen j., mack rl (eds.), usability inspection methods.
- M. Niu, L. Yu, S. Tian, X. Wang, and Q. Zhang. 2020. Personal-bullying detection based on multi-attention and cognitive feature. *Automatic Control and Computer Sciences*, 54:52–61.
- Annalaura Nocentini, Juan Calmaestra, Anja Schultze-Krumbholz, Herbert Scheithauer, Rosario Ortega, and Ersilia Menesini. 2010. Cyberbullying: Labels, behaviours and definition in three european countries. *Australian Journal of Guidance and Counselling*, 20:129.
- Donald A Norman. 1986. Cognitive engineering. *User centered system design*, 31:61.
- Ofcom Research. 2019. Online nation.

- Sei-Youen Oh. 2019. Cyberbullying response system on sns. *International Journal of Recent Technology and Engineering*, 8:241 – 244.
- Nuno Oliveira, Paulo Cortez, and Nelson Areal. 2013. On the predictability of stock market behavior using stocktwits sentiment and posting volume. pages 355–365. Springer.
- Dan Olweus. 2012. Cyberbullying: An overrated phenomenon? *European Journal of Developmental Psychology*, 9:520–538.
- Dan Olweus and Susan P. Limber. 2018. Some problems with cyberbullying research.
- Juri Opitz and Sebastian Burst. 2019. Macro f1 and macro f1. *arXiv*.
- Silviu Oprea and Walid Magdy. 2019. isarcasm: A dataset of intended sarcasm. *arXiv*.
- Rosario Ortega-Ruiz, Rosario Del Rey, and José A. Casas. 2012. Knowing, building and living together on internet and social networks: The conred cyberbullying prevention program - proquest. *International Journal of Conflict and Violence*, 6:302 – 312.
- Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. *arXiv preprint arXiv:1908.11049*.
- Sara Pabian and Heidi Vandebosch. 2016. (cyber)bullying perpetration as an impulsive, angry reaction following (cyber)bullying victimisation?
- Claus Pahl and Pooyan Jamshidi. 2016. Microservices: A systematic mapping study. pages 137–146.
- Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.
- Kartikey Pant and Tanvi Dadu. 2020. Sarcasm detection using context separators in online discourse.
- Lucy Pasha-Robinson. 2012. Teenager killed herself in front of parents after 'relentless' cyber bullying — the independent.
- Sayanta Paul and Sriparna Saha. 2020. Cyberbert: Bert for cyberbullying identification: Bert for cyberbullying identification. volume 1, page 3. Springer Science and Business Media Deutschland GmbH.

- Hanchuan Peng, Fuhui Long, and Chris Ding. 2005. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1226–1238.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. pages 1532–1543.
- Pew Research Center. 2018. A majority of teens have experienced some form of cyberbullying.
- Jon Porter. 2019. Instagram to start warning users before they post ‘potentially offensive’ captions.
- Rolandos Alexandros Potamias, Georgios Siolas, and Andreas Georgios Stafylopatis. 2020. A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, 32:17309–17320.
- N. Potha, M. Maragoudakis, and D. Lyras. 2016. A biology-inspired, data mining framework for extracting patterns in sexual cyberbullying data. *Knowledge-Based Systems*, 96:134–155.
- Nektaria Potha and Manolis Maragoudakis. 2015. Cyberbullying detection using time series modeling. volume 2015-January, pages 373–382. IEEE Computer Society.
- Ankit Pradhan, Venu Madhav Yatam, and Padmalochan Bera. 2020. Self-attention for cyberbullying detection. Institute of Electrical and Electronics Engineers Inc.
- Michal Ptaszynski, Juuso Kalevi Kristian Eronen, and Fumito Masui. 2017. Learning deep on cyberbullying is always better than brute force. pages 19–25.
- Perla Janeth Castro Pérez, Christian Javier Lucero Valdez, María de Guadalupe Cota Ortiz, Juan Pablo Soto Barrera, and Pedro Flores Pérez. 2012. Misaac: Instant messaging tool for ciberbullying detection. page 1. The Steering Committee of The World Congress in Computer Science, Computer
- Syarifah Qonitatulhaq, Novita Astin, and Widi Sarinastiti. 2019. Creative media for cyberbullying education. pages 622–627. IEEE.
- Evani Radiya-Dixit and Xin Wang. 2020. How fine can fine-tuning be? learning efficient language models.

- Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2020. Bullyalert- a mobile application for adaptive cyberbullying detection. volume 341, pages 3–15. Springer Science and Business Media Deutschland GmbH.
- Rahat Ibn Rafiq, Homa Hosseinmardi, Richard Han, Qin Lv, and Shivakant Mishra. 2018. Scalable and timely detection of cyberbullying in online social networks. pages 1738–1747. Association for Computing Machinery.
- Rahat Ibn Rafiq, Homa Hosseinmardi, Richard Han, Qin Lv, Shivakant Mishra, and Sabrina Arredondo Mattson. 2015. Careful what you share in six seconds: Detecting cyberbullying instances in vine. pages 617–622. Association for Computing Machinery, Inc.
- Elaheh Raisi and Bert Huang. 2018a. Weakly supervised cyberbullying detection using co-trained ensembles of embedding models. pages 479–486. Institute of Electrical and Electronics Engineers Inc.
- Elaheh Raisi and Bert Huang. 2018b. Weakly supervised cyberbullying detection with participant-vocabulary consistency. *Social Network Analysis and Mining*, 8.
- Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter:a behavioral modeling approach. pages 97–106. Association for Computing Machinery, Inc.
- Tazeek Bin Abdur Rakib and Lay-Ki Soon. 2018. Using the reddit corpus for cyberbully detection. pages 180–189. Springer.
- Kuldeep Randhawa, Chu Kiong Loo, Manjeevan Seera, Chee Peng Lim, and Asoke K. Nandi. 2018. Credit card fraud detection using adaboost and majority voting. *IEEE Access*, 6:14277–14284.
- Steve Ranger. 2016. Google accused of abusing android smartphone dominance to stifle competition.
- Janet C Read, Daniel Fitton, and Matthew Horton. 2014. Giving ideas an equal chance: Inclusion and representation in participatory design with children. *dl.acm.org*, pages 105–114.

- Roberta Renati, Carlo Berrone, and Maria Assunta Zanetti. 2012. Morally disengaged and unempathic: Do cyberbullies fit these definitions? an exploratory study. *Cyberpsychology, Behavior, and Social Networking*, 15:391–398.
- Mohammadreza Rezvan, Saeedeh Shekarpour, Lakshika Balasuriya, Krishnaprasad Thirunarayan, Valerie L Shalin, and Amit Sheth. 2018. A quality type-aware annotated corpus and lexicon for harassment research. pages 33–36.
- Irina Rish. 2001. An empirical study of the naive bayes classifier. volume 3, pages 41–46.
- Walisa Romsaiyud, Kodchakorn Na Nakornphanom, Pimpaka Prasertsilp, Piyaporn Nurarak, and Pirom Konglerd. 2017. Automated cyberbullying detection using clustering appearance patterns. pages 242–247. Institute of Electrical and Electronics Engineers Inc.
- Hugo Rosa, David Matos, Ricardo Ribeiro, Luisa Coheur, and João P. Carvalho. 2018. A 'deeper' look at detecting cyberbullying in social networks. volume 2018-July. Institute of Electrical and Electronics Engineers Inc.
- Cornelia M. Ruland, Justin Starren, and Torun M. Vatne. 2008. Participatory design with children in the development of a support system for patient-centered care in pediatric oncology. *Journal of Biomedical Informatics*, 41:624–635.
- Baidya Nath Saha and Apurbalal Senapati. Cit kokrajhar team: Lstm based deep rnn architecture for hate speech and offensive content (hasoc) identification in indo-european languages.
- Hassan Saif, Yulan He, and Harith Alani. 2012. Semantic sentiment analysis of twitter. pages 508–524. Springer.
- Semiu Salawu, Yulan He, and Joanna Lumsden. 2017. Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing*, 11(1):3–24.
- Huascar Sanchez and Shreyas Kumar. 2011. Twitter bullying detection. *ser. NSDI*, 12:15.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

- Hagit Sasson and Gustavo Mesch. 2014. Parental mediation, peer norms and risky online behavior among adolescents. *Computers in Human Behavior*, 33:32–38.
- Jan Schilling. 2006. On the pragmatics of qualitative assessment european. *Journal of Psychological Assessment*, 22:28–37.
- Jürgen Schmidhuber. 2015. Deep learning in neural networks: An overview.
- Douglas Schuler and Aki Namioka. 1993. *Participatory design: Principles and practices*. CRC Press.
- Suranga Seneviratne, Aruna Seneviratne, Prasant Mohapatra, and Anirban Mahanti. 2015. Your installed apps reveal your gender and more! *ACM SIGMOBILE Mobile Computing and Communications Review*, 18:55–61.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. volume 3, pages 1715–1725. Association for Computational Linguistics (ACL).
- Stephen M. Serra and H. S. Venter. 2011. Mobile cyber-bullying: A proposal for a pre-emptive approach to risk mitigation by employing digital forensic readiness.
- J. I. Sheeba and K. Vivekanandan. 2013. Low frequency keyword extraction with sentiment classification and cyberbully detection using fuzzy logic technique. IEEE Computer Society.
- Atanu Shome, Md Mizanur Rahman, Sriram Chellappan, and A. B.M. Alim Al Islam. 2019. A generalized mechanism beyond nlp for real-time detection of cyber abuse through facial expression analytics. pages 348–357. Association for Computing Machinery.
- Marko Robnik-Šikonja and Igor Kononenko. 2003. Theoretical and empirical analysis of relieff and rrelieff.
- Yasin N. Silva, Deborah L. Hall, and Christopher Rich. 2018. Bullyblocker: toward an interdisciplinary approach to identify cyberbullying. *Social Network Analysis and Mining*, 8.
- Yasin N. Silva, Christopher Rich, and Deborah Hall. 2016. Bullyblocker: Towards the identification of cyberbullying in social networking sites. pages 1377–1379. Institute of Electrical and Electronics Engineers Inc.

- Vivek K. Singh, Souvick Ghosh, and Christin Jose. 2017. Toward multimodal cyberbullying detection. volume Part F127655, pages 2090–2099. Association for Computing Machinery.
- Robert Slonje and Peter K. Smith. 2008. Cyberbullying: Another main type of bullying? *Scandinavian Journal of Psychology*, 49:147–154.
- Donnavieve N Smith and Xiaoye Chen. 2018. Brand experience, flow and brand app loyalty: Examining consumer decision making within branded mobile apps. *Marketing Management Journal*, 28.
- Peter K. Smith. 2012. Cyberbullying: Challenges and opportunities for a research program—a response to olweus (2012). *European Journal of Developmental Psychology*, 9:553–558.
- Peter K. Smith. 2015. The nature of cyberbullying and what we can do about it. *Journal of Research in Special Educational Needs*, 15:176–184.
- Peter K. Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. 2008. Cyberbullying: Its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 49:376–385.
- Patricia Snell and Elizabeth Englander. 2010. Cyberbullying victimization and behaviors among girls: Applying research findings in the field. *Journal of Social Sciences*, pages 510 – 514.
- Hajung Sohn and Hyunju Lee. 2019. Mc-bert4hate: Hate speech detection using multi-channel bert for different languages and translations. volume 2019-November, pages 551–559. IEEE Computer Society.
- Mona E. Solberg and Dan Olweus. 2003. Prevalence estimation of school bullying with the olweus bully/victim questionnaire. *Aggressive Behavior*, 29:239–268.
- Joan Solsman. 2019. Spotify: Apple’s app store abuses its power to ‘stifle’ rivals.
- Devin Soni and Vivek Singh. 2018. Time reveals all wounds: Modeling temporal dynamics of cyberbullying sessions. pages 684–687. AAAI press.
- Sara Owsley Sood, Judd Antin, and Elizabeth Churchill. 2012a. Using crowdsourcing to improve profanity detection.

- Sara Owsley Sood, Elizabeth F. Churchill, and Judd Antin. 2012b. Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology*, 63:270–285.
- A. Squicciarini, S. Rajtmajer, Y. Liu, and C. Griffin. 2015. Identification and characterization of cyberbullying dynamics in an online social network. pages 280–285. Association for Computing Machinery, Inc.
- StackOverflow. 2008. performance - what is considered a good response time for a dynamic, personalized web application?
- Statista. 2019. Global mobile social penetration rate 2019, by region.
- Statista. 2020a. Facebook: users by age and gender.
- Statista. 2020b. Global twitter user age distribution 2020.
- Statista. 2020c. Instagram: age and gender demographics.
- Nick Statt. 2020. Twitter tests a warning message that tells users to rethink offensive replies.
- David W Stewart and Prem N Shamdasani. 2014. *Focus groups: Theory and practice*, volume 20. Sage publications.
- Fabio Sticca and Sonja Perren. 2013. Is cyberbullying worse than traditional bullying? examining the differential roles of medium, publicity, and anonymity for the perceived severity of bullying. *Journal of youth and adolescence*, 42:739–750.
- Lara Stocchi, Nina Michaelidou, and Milena Micevski. 2019. Drivers and outcomes of branded mobile app usage intention. *Journal of Product and Brand Management*, 28:28–49.
- Kaveri Subrahmanyam and Patricia Greenfield. 2008. Online communication and adolescent relationships.
- Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. Mobilebert: a compact task-agnostic bert for resource-limited devices. *arXiv*.
- Shan Suthaharan. 2016. Support vector machine. In *Machine learning models and algorithms for big data classification*, Integrated Series in Information Systems, pages 207–235. Springer.

- David Sweet, David Faure, Kurt Granroth, Daniel Marjamäki, Ralf Nolden, Charles Samuels, Espen Sand, Cristian Tibirna, and Stefan Westerfeld. 2001. Kde 2.0 development.
- Sajedul Talukder and Bogdan Carbutar. 2018. Abusniff: Automatic detection and defenses against abusive facebook friends.
- Mircea-Adrian Tanase, George-Eduard Zaharia, Dumitru-Clementin Cercel, and Mihai Dascalu. 2020. Detecting aggressiveness in mexican spanish social media content by fine-tuning transformer-based models.
- Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. pages 1555–1565.
- I. Hsien Ting, Wun Sheng Liou, Dario Liberona, Shyue Liang Wang, and Giovanni Mauricio Tarazona Bermudez. 2017. Towards the detection of cyberbullying based on social network mining techniques. volume 2018-January, pages 1–2. Institute of Electrical and Electronics Engineers Inc.
- Robert S. Tokunaga. 2010. Following you home from school: A critical review and synthesis of research on cyberbullying victimization.
- Çiğdem Topcu and Özgür Erdur-Baker. 2012. Affective and cognitive empathy as mediators of gender differences in cyber and traditional bullying. *School Psychology International*, 33:550–561.
- Sarit Tresser. 2017. Personalization of virtual games for children with cerebral palsy. pages 209–212. Association for Computing Machinery.
- Jatin Karthik Tripathy, S. Sibi Chakkaravarthy, Suresh Chandra Satapathy, Madhulika Sahoo, and V. Vaidehi. 2020. Albert-based fine-tuning model for cyberbullying analysis. volume 1, page 3. Springer Science and Business Media Deutschland GmbH.
- Estera Twardowska-Staszek, Izabela Zych, and Rosario Ortega-Ruiz. 2018. Bullying and cyberbullying in polish elementary and middle schools: Validation of questionnaires and nature of the phenomena. *Children and Youth Services Review*, 95:217–225.

- Heidi Vandebosch and Katrien Van Cleemput. 2008. Defining cyberbullying: A qualitative research into the perceptions of youngsters. *Cyberpsychology and Behavior*, 11:499–503.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. pages 5998–6008.
- Marloes D. A. van Verseveld, Minne Fekkes, Ruben G. Fekkink, and Ron J. Oostdam. 2020. Teachers' experiences with difficult bullying situations in the school: An explorative study. *The Journal of Early Adolescence*, page 027243162093919.
- Nishant Vishwamitra, Xiang Zhang, Jonathan Tong, Hongxin Hu, Feng Luo, Robin Kowalski, and Joseph Mazer. 2017. Mcdefender: Toward effective cyberbullying defense in mobile online social networks. pages 37–42. Association for Computing Machinery, Inc.
- Miriam Walker, Leila Takayama, and James A Landay. 2002. High-fidelity or low-fidelity, paper or computer? choosing attributes when testing web prototypes. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 46, pages 661–665. SAGE Publications Sage CA: Los Angeles, CA.
- Annalu Waller, Victoria Franklin, Claudia Pagliari, and Stephen Greene. 2006. Participatory design of a text message scheduling system to support young people with diabetes. *Health Informatics Journal*, 12:304–318.
- Greg Walsh. 2009. Wii can do it: Using co-design for creating an instructional game. pages 4693–4698. ACM Press.
- Greg Walsh, Alison Druin, Mona Leigh Guha, Elizabeth Foss, Evan Golub, Leshell Hatley, Elizabeth Bonsignore, and Sonia Franckel. 2010. Layered elaboration: A new technique for co-design with children. volume 2, pages 1237–1240. ACM Press.
- Jing Wang, Ronald J. Iannotti, and Tonja R. Nansel. 2009. School bullying among adolescents in the united states: Physical, verbal, relational, and cyber. *Journal of Adolescent Health*, 45:368–375.
- Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2014. Cursing in english on twitter. pages 415–425.

- Tom Warren. 2020. Apple faces another eu antitrust complaint as app store pressure grows.
- Zeeraq Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. pages 88–93.
- D. Y. Weider, M. Gole, Nishanth Prabhuswamy, Sowmya Prakash, and Vidya Gowdru Shankaramurthy. 2016. An approach to design and analyze the framework for preventing cyberbullying. pages 864–867. Institute of Electrical and Electronics Engineers Inc.
- Sue Wilkinson. 2011. Analysing focus group data. *Qualitative research*, 3:168–184.
- N. Willard. 2005. An educator’s guide to cyberbullying and cyberbthreats: Responding to the challenge of online social aggression, threats, and distress.
- Kirk R. Williams and Nancy G. Guerra. 2007. Prevalence and predictors of internet bullying. *Journal of Adolescent Health*, 41.
- Laurie Williams and Robert R Kessler. 2003. *Pair programming illuminated*. Addison-Wesley Professional.
- Laurie Williams, Robert R. Kessler, Ward Cunningham, and Ron Jeffries. 2000. Strengthening the case for pair programming. *IEEE Software*, 17:19–25.
- V. Skye Wingate, Jessy A. Minney, and Rosanna E. Guadagno. 2013. Sticks and stones may break your bones, but words will always hurt you: A review of cyberbullying.
- Michelle F. Wright. 2016. Cybervictimization and substance use among adolescents: The moderation of perceived social support. *Journal of Social Work Practice in the Addictions*, 16:93–112.
- Jiale Wu, Mi Wen, Rongxing Lu, Beibei Li, and Jinguo Li. 2020. Toward efficient and effective bullying detection in online social network. *Peer-to-Peer Networking and Applications*.
- Jaideep Yadav, Devesh Kumar, and Dheeraj Chauhan. 2020. Cyberbullying detection using pre-trained bert model. pages 1096–1100. Institute of Electrical and Electronics Engineers Inc.

- Xiaohui Yang, Zhenhong Wang, Huan Chen, and Danni Liu. 2018. Cyberbullying perpetration among chinese adolescents: The role of interparental conflict, moral disengagement, and moral identity. *Children and Youth Services Review*, 86:256–263.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding.
- Mariya Yao. 2015. Three models of pair design. pair programming is an agile software. . . — by mariya yao — ux collective.
- Mengfan Yao, Charalampos Chelmiss, and Daphney Stavroula Zois. 2018. Cyberbullying detection on instagram with optimal online feature selection. pages 401–408. Institute of Electrical and Electronics Engineers Inc.
- Mengfan Yao, Charalampos Chelmiss, and Daphney Stavroula Zois. 2019. Cyberbullying ends here: Towards robust detection of cyberbullying in social media. pages 3427–3433. Association for Computing Machinery, Inc.
- Michele L. Ybarra and Kimberly J. Mitchell. 2004. Online aggressor/targets, aggressors, and targets: a comparison of associated youth characteristics. *Journal of Child Psychology and Psychiatry*, 45:1308–1316.
- Michele L. Ybarra and Kimberly J. Mitchell. 2008. How risky are social networking sites? a comparison of places online where youth sexual solicitation and harassment occurs. *Pediatrics*, 121.
- Michele L. Ybarra, Kimberly J. Mitchell, Neal A. Palmer, and Sari L. Reisner. 2015. Online social support as a buffer against online and offline peer and sexual victimization among u.s. lgbt and non-lgbt youth. *Child Abuse and Neglect*, 39:123–136.
- Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, and Lynne Edwards. 2009. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB*, 2:1–7.
- Kelly Young and Catherine Govender. 2018. A comparison of gender, age, grade, and experiences of authoritarian parenting amongst traditional and cyberbullying perpetrators. *South African Journal of Education*, 38.

- Wael Shaher Mohammed Yousef and Al Bellamy. 2015. The impact of cyberbullying on the self-esteem and academic functioning of arab american middle and high school students. *Electronic Journal of Research in Educational Psychology*, 13:1696–2095.
- Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:1415–1420.
- Anman Zhang, Bohan Li, Shuo Wan, and Kai Wang. 2019. Cyberbullying detection with birnn and attention mechanism. volume 294 LNCIST, pages 623–635. Springer.
- Xiang Zhang, Jonathan Tong, Nishant Vishwamitra, Elizabeth Whittaker, Joseph P Mazer, Robin Kowalski, Hongxin Hu, Feng Luo, Jamie Macbeth, and Edward Dillon. 2016. Cyberbullying detection with a pronunciation based convolutional neural network. pages 740–745. IEEE.
- Rui Zhao, Anna Zhou, and Kezhi Mao. 2016. Automatic detection of cyberbullying on social networks based on bullying features. volume 04-07-January-2016. Association for Computing Machinery.
- Zehua Zhao, Min Gao, Fengji Luo, Yi Zhang, and Qingyu Xiong. 2020. Lshwe: Improving similarity-based word embedding with locality sensitive hashing for cyberbullying detection. Institute of Electrical and Electronics Engineers Inc.
- Haoti Zhong, David J. Miller, and Anna Squicciarini. 2019. Flexible inference for cyberbully incident detection. volume 11053 LNAI, pages 356–371. Springer Verlag.
- Zongkui Zhou, Hanying Tang, Yuan Tian, Hua Wei, Fengjuan Zhang, and Chelsey M. Morrison. 2013. Cyberbullying and its risk factors among chinese high school students. *School Psychology International*, 34:630–647.
- Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. pages 19–27.
- Elizaveta Zinovyeva, Wolfgang Karl Härdle, and Stefan Lessmann. 2020. Antisocial online behavior detection using deep learning. *Decision Support Systems*, 138:113362.

- Daphney Stavroula Zois, Angeliki Kapodistria, Mengfan Yao, and Charalampos Chelmiss. 2018. Optimal online cyberbullying detection. volume 2018-April, pages 2017–2021. Institute of Electrical and Electronics Engineers Inc.
- Ágnes Zsila, Róbert Urbán, Mark D. Griffiths, and Zsolt Demetrovics. 2019. Gender differences in the association between cyberbullying victimization and perpetration: The role of anger rumination and traditional bullying experiences. *International Journal of Mental Health and Addiction*, 17:1252–1267.
- Tomczyk Łukasz and Włoch Anna. 2019. Cyberbullying in the light of challenges of school-based prevention. *International Journal of Cognitive Research in Science, Engineering and Education*, 7:13–26.
- Mustafa Şahin. 2012. The relationship between the cyberbullying/cybervictimization and loneliness among adolescents. *Children and Youth Services Review*, 34:834–837.

APPENDIX A.1: Messages used to evaluate Cyberbullying Prevention Mobile Apps.

Obvious Abusive Messages
Ha ha u r so fat Go kill yourself You are such a loser I hate you Me and my guys gonna smash your head in Dude how gay r you lol Shut up u fag U r such a retard U r so annoying I will whoop
Less Obvious Abusive Messages
You are not the smartest bulb, are you? U look preggers lol U r sooooo faaaaat Have you ever thought of a face transplant U r nothing but an empty headed daft punk Have some class clearly wit is a skill you r yet to master You have got a face on you like sour lemons You are quite cute until you opened your mouth U look so pretty with makeup on

APPENDIX B.1: Study's Ethics Approval.

From: Christopher Buckingham <c.d.buckingham@aston.ac.uk>
Sent: 27 September 2016 16:18
To: Salawu, Semiu (Research Student) <salawusd@aston.ac.uk>
Cc: Lumsden, Joanna (Jo) <J.LUMSDEN@aston.ac.uk>; Leigh, John <j.leigh@aston.ac.uk>; Doughty, Susan <s.m.doughty@aston.ac.uk>
Subject: Ethics Application 952

Dear Semiu,

Thank you for your resubmission, which the ethics committee has now reviewed. In general, you have made considerable improvements and we are happy for the research to go ahead on the understanding that it conforms with the statements you have made in the ethics submission and accommodates our original feedback. However, we would also like to clarify the following conditions for the research:

- 1) There will be NO videos of any student participants for any stage of the project. Videos of the designs are fine but not of the people doing them.
- 2) The Participatory Design must NOT discuss or explore in any way the attitudes or experiences of students relating to bullying (for the reasons given in the original feedback).
- 3) Although you have improved the explanation of who will be selected for the various stages, we still think it could be better. To stop anyone feeling they have been left out, please put in a statement along the following lines: "if we have more volunteers than places for each stage of the research, we will draw names out by chance so apologies in advance if your name is not one of them". Then they know random selection is involved and that they have not been rejected.
- 4) There must be an explicit statement from the school counselling service that it will provide support for any students who need it as a result of participating in the project.

Good luck with the project and stay in touch if you need any further help or advice.

Best wishes,

Chris
EAS Ethics Chair

Christopher Buckingham TEL: 0121-204-3450
Senior Lecturer email: C.D.Buckingham@aston.ac.uk
Computer Science, Aston University Fax +44 (0)121 204 3681
Aston Triangle, Birmingham B4 7ET

APPENDIX B.2: Invitation Email for Adult Participants

Dear [*name*],

We would like to invite you to participate in a research project entitled ***Detection and Prevention of Cyberbullying in Social Networks***.

The purpose of this research project is to understand how bullying is perpetrated on social media and how best to detect and prevent it. A key part of this research is developing a new mobile app capable of detecting various forms of cyberbullying and taking actions on behalf of the user to reduce them.

Should you choose to participate, you will be asked to take part in the following activities:

- Focus Groups

The focus group sessions will involve discussing your views on cyberbullying and how cyberbullying can be prevented with the researcher and other participants. This research will benefit young people being bullied across the world, and by taking part, you will have contributed to advancing the research in cyberbullying prevention. Each session will last up to two hours, and it is envisaged that there will be 3 – 4 sessions. All sessions will be audio recorded. As token appreciation of your participation in this research, you will be entitled to a £10 Amazon voucher per each attended session.

If you want to be part of this research, please complete the attached consent form and email back to the researcher at **salawusd@aston.ac.uk**. A Participant Information Sheet is also attached, and this provides additional information about the research.

If you have any pertinent questions about your rights as a research participant, please contact the Aston University Research Ethics Committee via the details available at the link below:

<https://www.ethics.aston.ac.uk/content/committee-officers>

If you have any questions, please feel free to contact me at salawusd@aston.ac.uk.

Thank you,
Semiu Salawu, *PhD Researcher*.



EVALUATION OF BULLSTOP: A MOBILE APP FOR THE DETECTION AND PREVENTION OF CYBERBULLYING ON SOCIAL MEDIA

Participant Information Sheet

Invitation

We would like to invite you to take part in a research study.

Before you decide if you would like to participate, take time to read the following information carefully and, if you wish, discuss it with others such as your family, friends or colleagues.

Please ask a member of the research team, whose contact details can be found at the end of this information sheet, if there is anything that is not clear or if you would like more information before you make your decision.

What is the purpose of this study?

You are being asked to participate in a study to evaluate the usability and usefulness of a mobile app designed to detect and combat cyberbullying on social media. The app uses novel computing technology to detect various forms of cyberbullying and take appropriate actions (such as deleting offensive messages and blocking cyberbullies and trolls) on behalf of the user. The target audience for this mobile application will be young people in the UK aged 11 – 17 years. We hope that this app will help protect vulnerable people, especially children, from the damage caused by cyberbullying.

Why have I been chosen?

You are being invited to take part in this study because as a [*profession*], we believe your input will be very valuable in assessing the mobile app from the perspective of responsible adults in the lives of the young people targeted by the app.

What will happen to me if I take part?

You will be asked to evaluate the app by using an Android smartphone to complete a series of tasks within the app. After using the app to complete the tasks, you will be asked to discuss your opinions of the app with the researcher in an interview. This interview will be audio recorded so that we can maintain an accurate record of what you thought about the app.

To protect your identity, you will be provided with a dummy account to login and use the app, and you will also be provided with a dummy Twitter account to which to connect the app. In other words, none of your personal contact information or your social media activity will be visible to the researchers or recorded as part of this study.

As the app is designed to be intuitive and simple to use, no special skills are required to use the app. It is designed for the same level of app competence as any social media app. You can however ask questions or request assistance from the researcher if required at any time. In addition, the app includes a tutorial on how to use the app which you can refer to whenever you wish.

The app is designed to record and track user actions such as screen navigation and taps. Keystrokes are however NOT recorded and so anything you type into the app is not recorded.

The entire session is expected to last between 60 – 90 minutes.

Do I have to take part?

No. It is up to you to decide whether or not you wish to take part. If you do decide to participate, you will be asked to sign and date a consent form. You will still be free to withdraw from the study at any time without giving a reason.

Will my taking part in this study be kept confidential?

Yes. A code will be attached to all the data you provide to maintain confidentiality.

Your personal data (name and contact details) will only be used if the researchers need to contact you to arrange study visits or collect data by phone. Analysis of your data will be undertaken using coded data.

The data we collect will be stored in a secure document store (paper records) or electronically on a secure encrypted mobile device, password protected computer server or secure cloud storage device.

To ensure the quality of the research, Aston University may need to access your data to check that the data has been recorded accurately. If this is required, your personal data will be treated as confidential by the individuals accessing your data.

How will the conversations that take place during the interview be recorded and the information I provide managed?

With your permission we will audio record the interview and take notes.

The recording will be typed into a document (transcribed) by the researcher. This process will involve removing any information which could be used to identify individuals e.g., names, locations, etc.

Audio recordings will be destroyed as soon as the transcripts have been checked for accuracy. We will ensure that anything you have told us that is included in the reporting of the study is anonymous.

You of course are free not to answer any questions that are asked without giving a reason.

What are the possible benefits of taking part?

While there are no direct benefits to you of taking part in this study, the data gained will contribute to knowledge that will allow us to make this app as good as it can be and, in future, design other applications to mitigate and prevent cyberbullying.

What are the possible risks and burdens of taking part?

There are minimal risks associated with participating in this study beyond that of normal everyday usage of social media. The app (BullStop) is targeting a sensitive issue and, in order to allow participants to properly evaluate its use, the app has been pre-populated with 'fake' bullying content. A small portion of these offensive messages may contain one or more of the following profane words – damn, hell, sh*t, fu*k. None of the messages include racist, homophobic, transphobic, sexist and offensive content about age, weight, physical appearance and religion.

Whilst the fake content will be carefully selected to avoid being unduly offensive, it is recognised that it could be offensive for some and/or could trigger memories of personal bullying for some

participants. If any of the bullying messages make you feel uncomfortable then you should let the researcher know immediately; similarly, if the researcher observes you exhibiting behaviour indicative of psychological discomfort, he will suspend your involvement until convinced that you are able to and wish to continue. If participation in the study causes distress, you will be directed to a range of help resources (links to which are also embedded in the app itself).

You can at any point stop your participation in the study. This does not affect any payments to which you are entitled.

What will happen to the results of the study?

The results of this study may be published/presented in scientific journals and/or presented at conferences. If the results of the study are published, your identity will remain confidential.

A lay summary of the results of the study will be available for participants when the study has been completed and the researchers will ask if you would like to receive a copy.

The results of the study will also be used in the PhD thesis of Semiu Salawu.

Expenses and payments

You will be entitled to £10 worth of Amazon vouchers for taking part in this study. No expenses payments will be made.

Who is funding the research?

This research is self-funded by the researcher (Semiu Salawu).

Who is organising this study and acting as data controller for the study?

Aston University is organising the study and acting as data controller for the study. You can find out more about how we use your information in Appendix A.

Who has reviewed the study?

This study was given a favourable ethical opinion by Aston University Research Ethics Committee.

What if I have a concern about my participation in the study?

If you have any concerns about your participation in this study, please speak to the research team and they will do their best to answer your questions. Contact details can be found at the end of this information sheet.

If the research team are unable to address your concerns or you wish to make a complaint about how the study is being conducted, you should contact the Aston University Research Integrity Office at research_governance@aston.ac.uk or telephone 0121 204 3000.

Research Team

If you have any questions, you may contact the PhD student researcher or his supervisor at:

Semiu Salawu (PhD Student), School of Engineering & Applied Science, Aston University, e-Mail: salawusd@aston.ac.uk.

Dr Jo Lumsden (Supervisor), School of Engineering & Applied Science, Aston University, e-Mail: j.lumsden@aston.ac.uk. Tel: 0121 204 3470

Thank you for taking time to read this information sheet. If you have any questions regarding the study, please don't hesitate to ask one of the research team.



Aston University takes its obligations under data and privacy law seriously and complies with the General Data Protection Regulation (“GDPR”) and the Data Protection Act 2018 (“DPA”).

Aston University is the sponsor for this study based in the United Kingdom. We will be using information from you in order to undertake this study. Aston University will process your personal data in order to register you as a participant and to manage your participation in the study. It will process your personal data on the grounds that it is necessary for the performance of a task carried out in the public interest (GDPR Article 6(1)(e)). Aston University may process special categories of data about you which includes details about your health. Aston University will process this data on the grounds that it is necessary for statistical or research purposes (GDPR Article 9(2)(j)). Aston University will keep identifiable information about you for 6 years after the study has finished.

Your rights to access, change or move your information are limited, as we need to manage your information in specific ways in order for the research to be reliable and accurate. If you withdraw from the study, we will keep the information about you that we have already obtained. To safeguard your rights, we will use the minimum personally identifiable information possible.

You can find out more about how we use your information at www.aston.ac.uk/dataprotection or by contacting our Data Protection Officer at dp_officer@aston.ac.uk.

If you wish to raise a complaint on how we have handled your personal data, you can contact our Data Protection Officer who will investigate the matter. If you are not satisfied with our response or believe we are processing your personal data in a way that is not lawful you can complain to the Information Commissioner’s Office (ICO).



WEST MIDLANDS POLICE RESEARCH APPLICATION

OFFICE USE ONLY:

Reference	
Decision & Date	

The Academic Research Team is responsible for monitoring all research undertaken within the force using force data, facilities, resources or staff and officer duty time.

This is to ensure force data is used according to our statutory obligations under Data Protection and to co-ordinate research activities, avoiding duplication and maximising the benefits to the organisation and the communities we serve.

This application form must be completed by any individual seeking to carry out a research project using Force data, systems or consulting with staff or officers. Applications are evaluated using several criteria including:-

- **Legal frameworks (Data Protection)**
- **Security of Information**
- **Value of the research**
- **Availability of data/information**
- **Ease of data abstraction/abstraction of resource from core duties**
- **Reputation and expertise of the researcher/research institution**

**If you experience any difficulties in completing this form please contact a member of the Research Team using email address:
academic_research@west-midlands.pnn.police.uk**

RESEARCH CANNOT BE COMMENCED UNTIL APPROVED BY WMP

Section 1 – Your Details

- **Your Details: (All Non WMP Employee Applicants To Complete)**

Name:	Semiu Salawu
Home Address:	30 Oval Drive, Wolverhampton, WV10 6AX
Landline Number:	01902651160
Mobile Number:	01902651160
Email Address:	salawusd@aston.ac.uk

- **Contact Details for WMP to use if different from above:**

Contact Address:	
Landline Number:	
Mobile Number:	
Email Address:	

Will the research form part of your:
(please delete one)

School or College Studies

Undergraduate Studies

Post Graduate Studies

- If this research will not form part of an educational qualification, please state your reasons for undertaking the research:

--

Please state the subject you are studying and qualification you are undertaking (e.g. Criminology PhD):

Subject: Computer Science

Qualification: PhD

Final Qualification Date: October 2019

Please give a) The name of your tutor b) The educational facility where you are studying c) The city/town location of the educational facility:

**Tutor's Name:
Dr. Jo Lumsden**

Educational Facility: Aston University

City/Town of Education Facility: Birmingham

Section 1 – Your Details

- Your Details (For WMP Employees to complete only)

**This section is for members of West Midlands Police (officers and staff).
If you are external to the organisation, please skip this section and continue your application in section 2.**

Name:	
Title/Rank:	
LPU or Dept:	
Internal Ext. Number:	

Is the research required for or part of an academic qualification? (<i>Please delete one</i>)
Yes
No

If 'Yes' what is the qualification?

- Are WMP supporting/funding this qualification? (Financial/Study Leave)? (<i>Please delete one</i>)
Yes
No

If 'Yes' what is the nature of the support you are receiving?

To progress your application to conduct research in West Midlands Police we require an endorsement from a member of your LPU/Department command team. This must be a senior member of staff of Superintending rank (or police staff equivalent). There is a section in the 'Your Research' part of the application for them to comment on the research you are proposing and it's value. If your application to conduct research is approved they may be asked to act as a sponsor and provide oversight on the appropriate use of force data, systems and resources used in the course of conducting this research.

Section 2 – About Your Research

What is the title of your proposed research project?

Detection and Prevention of Cyberbullying in Social Media

Please summarise the purpose or aims of your research

(Context for research including identification of key gaps, research question/s, aims and objectives, expected contribution to knowledge)

The key aims of the research are as follow:

- **To develop a mobile app that uses machine learning classifiers to detect cyberbullying and enable the app take preventive actions on behalf of the victims**
- **To gain insight into what UK adolescents understand by cyberbullying and what they perceive are the effects of cyberbullying**
- **To use participatory design methods to incorporate inputs from key stakeholders (adolescents, parents, teachers, law enforcement) into the design and development of cyberbullying prevention tools**

What are the envisaged benefits to WMP as an organisation or to policing?

The proposed mobile app is aimed at detecting cyberbullying directed to the user on social media websites. Once a cyberbullying incident has been detected, the app can then perform a range of configured actions. This can include automatically deleting or quarantining the message, blocking usage or forwarding the message to a nominated third party (e.g. a parent). The app can also be configured to record evidence that can be used by the Police to pursue prosecution. It is therefore important to get law enforcement's input and incorporate this into the app's design.

When do you need to start and complete your research?

(if appropriate, attach a Gantt Chart or similar outlining key stages)

Start Date: September 2016

Completion Date: April 2017

What methodology will be used to complete the research?

(Please delete as appropriate)

~~Analysis of force documents/policy~~

~~Analysis of Force Data~~

~~Surveys/Questionnaires/Interviews/Focus Groups with officers and/or staff~~

~~Pilot Studies~~

Other (please specify) : Participatory design. This is a design approach that democratically and actively involves stakeholders in the design process to help ensure the end result meets their needs and is likely to be accepted into use.

How will the data be analysed?

(Theoretical/critical framework, statistical techniques and/or analytical tools (e.g. content analysis))

The key outputs from the participatory design sessions will be a list of features accompanied by the relative importance of each feature to the participatory design group. This data does not need analysis, rather the final version of the developed app will be accessed by the participatory design group in terms of these documented features.

Has your research been granted ethical approval?

Yes

No

Please give details of the support from or access you require to West Midlands Police - including data, facilities, systems or resources (including staff and officers

to complete questionnaires, interviews etc). If you require force data, please be specific as you can about the exact nature of the data and append a separate sheet if necessary.

A maximum of 2 representatives of West Midlands Police to attend all participatory design sessions. It is currently envisaged that there will be 8 sessions in all, each lasting approximately 90 minutes. The sessions are to be held between September 2016 and April 2017 and will take place at Great Barr School, Aldridge Rd, Birmingham B44 8NU.

Please select one or more of the appropriate notified purposes below that state under which circumstances force data can be used:

(Please delete as appropriate)

The prevention and detection of crime

~~**Apprehension and prosecution of offenders**~~

~~**Protection of life and property**~~

~~**Maintenance of law and order**~~

~~**Vetting and licensing**~~

~~**Public safety**~~

~~**Rendering assistance to members of the public in line with force policy**~~

Are you seeking access to data where individuals are identified (personal data)
(Please delete one)

Yes

No

Who will be funding the research?

Self-funded.

Please give details of any experience you have in researching this area and your expertise. If you have authored published papers, please note them below or attach a publication list.

I started my research in this area in October 2013 and currently have a submitted paper titled “Approaches to Cyberbullying Detection” currently in review with ACM (Association for Computing Machinery) Transactions on Intelligent Systems and technology and a 2nd paper titled “A Survey of Cyberbullying Prevention Software ” undergoing supervisory review.

Please use the space below to add any other comments in support of your application or reference any other supporting documents you are submitting (e.g. – research brief, publication list)

FOR COMPLETION BY APPLICANT'S WMP LPU OR DEPARTMENT MANAGERS ONLY

This must be a senior member of staff of Superintending rank (or police staff equivalent).

Please comment on the value of this research proposal

Semiu is a very capable PhD student who is tackling a timely and necessary issue. His approach could see significant inroads to cyberbullying detection/prevention by using ubiquitous technology and, most importantly, by including target end users (teenagers) in the design of the software. His inclusion of law enforcement as stakeholders in the process is critical and valuable to ensuring the final solution meets needs from all perspectives and he would really value the input of members of your team in this respect.

Managers Name

Dr Joanna Lumsden

Rank/Role

Reader & Aston Interactive Media (AIM) Lab Manager

Signature

Date

24/5/2016

Section 3 – How We Use Your Research

By working in association with West Midlands Police, West Midlands Police will have an interest in the outcome of the research as well as the finished product.

Upon completion of your research, a copy of your finalised research paper must be sent to the Academic Research Team (academic_research@west-midlands.pnn.police.uk). This will be kept on record with the force, published onto our force internal intranet research page and potentially referenced to or utilised in the future.

You may also be asked to present your findings or recommendations from your paper to the West Midlands Police Command Team or at a force meeting or event.

By signing the below section you are confirming that you have read, understood and agree with the above requests:

Signed:	
Print Name:	
Date:	

Section 4 – Code of Ethics

If your application for research is successful, you will be considered as working in partnership with West Midlands Police, therefore, we will expect you to adhere to our Code of Ethics for the duration of your work with us.

The College of Policing developed the Code of Ethics on behalf of every member of the policing profession of England and Wales.

The main components of the Code are sets of principles and standards of professional behaviour.

The 9 policing principles are:

- **Accountability**
You are answerable for your decisions, actions and omissions
- **Fairness**
You treat people fairly
- **Honesty**
You are truthful and trustworthy
- **Integrity**
You always do the right thing
- **Leadership**
You lead by good example
- **Objectivity**
You make choices on evidence and your best professional judgement
- **Openness**
You are open and transparent in your actions and decisions
- **Respect**
You treat everyone with respect
- **Selflessness**
You act in the public interest

Should your conduct whilst working in conjunction with West Midlands Police be deemed as contravening our Code of Ethics, we shall review the complaint and take the necessary action. This may include revoking your vetting status and terminating the research project with you.

If you would like further information on our Code of Ethics, please do not hesitate to contact us.

Please sign the below box to indicate that you have read and understood the above and that you agree for the duration of your academic research project with West Midlands Police you shall adhere to the Code of Ethics.

Signed:	
Print Name:	
Date:	

Section 5 – What Happens Next?

Thank you for completing the research application form. Your application will be considered by a Commissioning Board which will deliberate the value, impact and data security issues arising from your request. The board meets monthly so it could be up to four weeks before your application is considered.

If you wish to check on the progress of your application, please contact the Research team using academic_research@west-midlands.pnn.police.uk or calling 101 and asking for Inspector Richard Harris.

The Board often need to re-contact applicants to clarify aspects of the proposed research, and this can add further delay, so please make sure:

- You have described accurately the potential benefits of the research
- You have described the type of support or access you require
- That your contact details and contact numbers are correct.

Successful applicants will be required to enter into further processes and complete additional documentation. These are likely to include vetting and a criminal records background check, a baseline security questionnaire and a data processing agreement. These processes and documentations may take a further four weeks to complete. In most circumstances they are required of the force and can not be circumvented. If the time scales for your research are particularly short you should consider alternative sources of information.

APPENDIX B.5: Consent Form for Adult Participants

CONSENT FORM

for

participation in the research study entitled:

Detection and Prevention of Cyberbullying in Social Networks

Name of Researcher:	Semiu Salawu
Participant ID:	

Please initial the boxes if you agree (DO NOT tick). One signed copy of the form will be for you and the other will be kept by the researcher.

1. I have read and understood the information sheet for the above project. I know that I can ask the researcher questions about the project.	<input type="checkbox"/>
2. I know that taking part is voluntary and I can stop taking part in the project at any time without giving reasons.	<input type="checkbox"/>
3. I understand that all information I give to the project will be kept private and my name and details will not be publicised.	<input type="checkbox"/>
4. I understand that my participation will be audio recorded.	<input type="checkbox"/>
5. I understand that, if I disclose information that would require disciplinary action under Aston University's normal policies and procedures, appropriate action will be taken by the University.	<input type="checkbox"/>
6. I agree for anonymised extracts from the audio recording to be used in any reports, publications or events where results from the study will be used.	<input type="checkbox"/>
7. I agree to take part in the above study.	<input type="checkbox"/>

Name (printed) and Signature

Date

Name of Researcher Obtaining Consent and Signature

Date

FOCUS GROUPS

Basic Information

Moderator: Semiu Salawu

Moderator email address: salawusd@aston.ac.uk

Participants:

- Parents.
- Teachers
- Mental Health Professionals
- Representatives of West Midlands Police.

Expected No. of Participants: 6- 8

Recording: Sessions will be **voice** recorded.

Duration: 60 – 90 minutes

No of Sessions: 2 - 3

Schedule: Schedule will be guided by the availability of participants.

Invitation to Participate

Parents, teachers, mental health professionals and law enforcement representatives will be invited via personal contact to participate. There are no inclusion criteria beyond their involvement in these roles.

Venue

The researcher will provide a comfortable space where participants can feel relaxed and that is spacious enough to allow for a circular seating arrangement.

Agenda

Introductions	–	10 mins
Discussion Part 1	–	20 - 30 mins
Break	–	5 - 10 mins
Discussion Part 2	–	20 - 30 mins
Closing Remarks	–	5 mins

Sessions Overview

Session 1:

Theme: *What is cyberbullying and how can it be prevented?*

Objective:

- *To gain an insight into what the group understand by cyberbullying and their opinions about it.*
- *To understand what they perceive are the effects of cyberbullying and the impact it has on adolescents.*

- *To generate prevention strategies from participants and compare these to current cyberbullying prevention strategies in use.*
- *To collectively assess the effectiveness of these prevention strategies and propose ways to improve the strategies.*

Key questions:

- *What do we understand by the term “Cyberbullying”?*
- *How important would you say preventing cyberbullying is to you?*
- *Is cyberbullying always malicious or just “teasing”?*
- *When does “teasing” becomes cyberbullying?*
- *Have you or do you know of anyone that has ever been cyberbullied?*
- *Have you or do you know of anyone that bullied someone?*
- *Please elaborate on these?*
- *What do we think are the effects of cyberbully?*
- *How can you tell that someone is being bullied online?*
- *What would you do if you discovered your child is being bullied online?*
- *What advice would you give to some that has been cyberbullied?*
- *What are the things we can do to prevent cyberbullying?*
- *If someone is being cyberbullied, what can they do to stop it?*
- *If someone we know (not our child) is being cyberbullied, would we do anything about it?*
- *What would we do?*
- *Are we aware of a situation where someone that was being bullied got the bully to stop?*
- *If yes what did the person do?*
- *Let’s create a list of suggested actions to reduce cyberbullying and rate these in order of preference?*

Session 2:

Theme: *Cyberbullying Prevention Software*

Objective:

- *To find out if participants currently use or will use anti-cyberbullying tools and their views about such tools.*
- *To collectively identify gaps in current anti-cyberbullying tools*
- *To present an overview of the proposed app to the group.*

Key questions:

- *Let’s review what we talked about in the previous session?*
- *Has anyone’s views on cyberbullying changed?*
- *How have these changed?*
- *Has anyone used a cyberbullying prevention tool before?*
- *If yes, which ones?*
- *If no, why?*
- *Do we think the anti-cyberbullying actions we developed last session can be automated?*
- *How will you like these actions to be implemented; by the service provider (e.g. Facebook), the school, as part of the device itself (pc or mobile) or as a software you can install and set up yourself.*

- *If implemented as a software, how confident are we that our children install and configure it by themselves?*
- *If we were to design an app to reduce cyberbullying, what are the things we will want in the app? Let's rate this in order of importance?*

Session 3:

Theme: *The Cyberbullying Prevention App*

Objective:

- *To review the proposed app with the group.*
- *To gather feedback about the proposed app*
- *To create a features "wish list" for the app*
- *To arrange the features "wish list" in order of importance*

Key questions:

- *If an app such as the one proposed was available on the app stores today, what would be our initial reactions to it?*
- *Would we encourage our children/students to use such an app?*
- *If no, what are the reasons?*
- *If the app detects a bullying message that has been sent to a child, what should it do? Should the message be automatically deleted? Should it be shown to the child with a warning? Should bad words in the message be removed/replaced before it is delivered? Should it be automatically forwarded to the parents?*
- *What about the senders of such messages? Should the app automatically block them or just block the messages? Or block them after a certain number of messages?*
- *Would you as a parent like the ability to remotely review and re-classify all messages flagged as bullying by the app?*
- *How important is such a feature to you*
- *Would you like the app to learn from this review and then base future decisions on your re-classification?*
- *How important is such a feature to you?*

Moderator's Script

Ensure consent form for all participants are duly completed and signed. Remind participants of confidentiality of content discussed in the group.

Introduction

(10 mins)

- Introduce self and provide study background
- Get participants to introduce themselves.

Discussion 1

(20 - 30 mins)

- Introduce the session's topic and objectives and start with the section's lead question.
- Gently progress discussion with additional questions.

Break

(5 - 10 mins)

Discussion 2

(20 - 30 mins)

- Quick recap of Discussion 1 then proceed with the section's lead question.
- Gently progress discussion with additional questions.

Closing Remarks

(05 mins)

- Quickly summarise point raised.
- Thank and inform group of the date and time of the next session.

APPENDIX B.7: Coding Tables for Focus Groups' Emergent Themes

Theme	Concerns about cyberbullying
Sample Quotes	Coded As
<p><i>"She didn't tell me about it for some time, and I had to prod her a bit, and it turns out it's even one of her close friends that I'm friendly with the mom, so I got the mum to have a word [...] I don't think they are friends anymore, but she stopped sending silly stuff to my girl anyway, and that's all I cared about, to be honest".</i></p> <p><i>"We get so many reports about stalking, abuse, nudes, all sorts [...], it's impossible to keep up. We don't have anywhere near the resources we need to handle all of it".</i></p> <p><i>"There is more about it on the news now, which I think is good. I saw on the news about a girl that was being bullied by her mates, they were sending nude pictures of her, but the funny thing is they weren't even her pictures. They just got some porn pic and cut her face on it, just like that. I felt that was just mean, and these are like 12, 13 year olds".</i></p> <p><i>"I get so worried and anxious with all the stuff online, sometimes I see things on some pages and [I'm] like wow".</i></p>	<ol style="list-style-type: none"> 1. Worried about the safety of children 2. Unable to keep up 3. Increased number of online attacks

Theme	Current strategies and solutions are in need of improvement
Sample Quotes	Coded As
<p><i>"I think the government should force them to do more".</i></p> <p><i>"I get freedom of speech, but some of these stuffs should be taken down immediately [...]"</i></p> <p><i>"I think my daughter reported someone that was sending these silly messages, but nothing really happened".</i></p>	<ol style="list-style-type: none"> 1. Social media companies need to do more 2. Government intervention is required
<p><i>"I wouldn't know what to do really, maybe tell my kids to ignore it".</i></p> <p><i>"Nothing really, I wasn't aware there were things I could use".</i></p> <p><i>"If it's really bad we sometimes recommend staying off social media for a while".</i></p> <p><i>"I went to school once to report a child for posting something on [child] Facebook and actually nothing really happened at first. I had to make a big deal, like go all crazy mum to get them to take it seriously".</i></p> <p><i>"I have searched online, and I found information on what to do, but they are not easily digestible. Something like a cheat sheet".</i></p>	<ol style="list-style-type: none"> 1. Unsure about what to do 2. Lack of information on available cyberbullying prevention tools 3. Schools are unwilling to get involved
<p><i>"I installed this app on [child]'s phone, it was meant to send me alerts when he uploads pictures on Instagram and things like when he sends messages and his location, but it kept crashing the phone, so we took it off. I'm sure he was very happy".</i></p> <p><i>"I used the Vodafone parental control on our broadband, just because I felt I had to do something. Like I can't just let them be browsing without any form of control. It blocks site and stuff".</i></p> <p><i>"I tell them to block, and I check their phones behind their backs".</i></p> <p><i>"I will like if it can tell me what my kids are doing online".</i></p>	<ol style="list-style-type: none"> 1. Ineffective prevention tools 2. Use of spyware 3. Constant monitoring

Theme	Encouraging positive behaviours and online safeguarding are key features
Sample Quotes	Coded As
<p><i>"I think like a safe browsing option will be good, so anything offensive is not shown to you when you are on Facebook and the likes".</i></p> <p><i>"I once read about having like a time out period from mobile phones and social media [...] I like that idea if you can put that in the app, it can just block out social media for like an hour or something".</i></p> <p><i>"I have an inspirational quotes app that I read in the morning. If the app can show something like that every day".</i></p>	<ol style="list-style-type: none"> 1. Making social media safer 2. Taking a break 3. Online safeguarding
<p><i>"[...] maybe it can rate children on how well they behave online".</i></p> <p><i>"It will be good if you can add links to some educational stuff about cyberbullying. It would be nice having all the information in one place".</i></p> <p><i>"It can include some videos on how to treat people when they are online".</i></p>	<ol style="list-style-type: none"> 1. Promote positive behaviours 2. Increase empathy 3. Encourage Reflection

Theme	Report and block online abusers
Sample Quotes	Coded As
<p><i>"If it can automatically block Internet trolls, I think that would be great".</i></p>	<ol style="list-style-type: none"> 1. Block abusive users
<p><i>"Can it report people to Facebook so they can be banned"?</i></p> <p><i>"[...] for really serious cases maybe report to the police or even just send them a text that you will be reported to the police".</i></p>	<ol style="list-style-type: none"> 1. Report bullies automatically 2. Involve the police

APPENDIX B.8: Ethics Approval for Engagement with First-Year Undergraduates

From: Christopher Buckingham <c.d.buckingham@aston.ac.uk>
Sent: 27 October 2017 12:48
To: Salawu, Semiu (Research Student) <salawusd@aston.ac.uk>
Cc: Lumsden, Joanna (Jo) <J.LUMSDEN@aston.ac.uk>; Leigh, John <j.leigh@aston.ac.uk>
Subject: Ethics resubmission 952

Dear Semiu,

We are happy with the extension of the proposal to first-year undergraduate students on the grounds that the same conditions apply to them as to the school students regarding the protocol. The one thing we would like to stress, because of the ambiguity in your statement, is that all participants should receive the SAME financial reward, whatever that may be.

Best wishes,

Chris
EAS Ethics Committee Chair

Christopher Buckingham TEL: 0121-204-3450, Room MB211N
Reader in Computer Science email: C.D.Buckingham@aston.ac.uk
Computer Science, Aston University Fax +44 (0)121 204 3681
Aston Triangle, Birmingham B4 7ET

Thousands of young people get bullied on social media everyday.

If you are a first year undergraduate, we would like your help in reducing cyberbullying.

Get up to £70 Amazon vouchers taking part in research investigating Cyberbullying on Social media.

Please contact salawusd@aston.ac.uk if you would like to take part.

**CYBERBULLYING IS
HAPPENING RIGHT NOW**

APPENDIX B.10: Invitation Email for First-Year Undergraduates

Dear student,

We would like to invite you to participate in a research project entitled ***Detection and Prevention of Cyberbullying in Social Networks***.

The purpose of this research project is to understand how bullying is perpetrated on social media and how best to detect and prevent it. A key part of this research is developing a new mobile app that is capable of detecting various forms of cyberbullying and taking actions on behalf of the user to reduce them.

Should you choose to participate, you will be asked to take part in the following activities:

- One on one interview
- Participatory design sessions

The interview will involve discussing your views on cyberbullying and how cyberbullying can be prevented with the researcher. This session will last up to an hour and will be audio recorded.

Participatory Design (PD) is an exciting design approach that democratically and actively involves potential users in the design process to help ensure the end result meets their needs and is likely to be accepted into use. Thus the participatory design sessions will involve working with other participants (all first-year undergraduates) and the researcher to design the key areas of the proposed mobile app. Each PD session will last up to 2 hours and 2 sessions are currently planned. All PD sessions will be audio and video recorded but please note that the video camera will not be directed at you, rather it will be pointed at a paper surface representing the app's user interface.

As token appreciation of your participation in this research, you will be entitled to up to **£70** Amazon vouchers (£10 for the interview and £20 for each of the 3 PD sessions). In addition, this research will benefit young people being bullied across the world and by taking part you will have contributed to advancing the research in cyberbullying prevention.

If you will like to be part of this research then please complete this [questionnaire](#). The questionnaire will allow us to get to know a little bit about you and your suitability to take part in the research. Please be aware that due to the need to ensure an appropriate distribution of participants, we will not be able to invite **everyone** that submit the questionnaire to progress to the interview and PD sessions stage but, at the end of this study, once we have developed the app, everyone will be given an opportunity to test the app and tell us what you think of it, regardless of whether they took part in the interview and PD sessions or not.

If you have any pertinent questions about your rights as a research participant, please contact the Aston University Research Ethics Committee via the details available at the link below:

<https://www.ethics.aston.ac.uk/content/committee-officers>

If you have any questions, please feel free to contact me at salawusd@aston.ac.uk.

Thank you,
Semiu Salawu, *PhD Researcher*.

Pre-Study Cyberbullying Questionnaire

1. What is your gender?
 - Male
 - Female
 - Prefer not to say

2. Are you a first-year undergraduate?
 - Male
 - Female
 - Prefer not to say

3. Please provide an email so we can contact you? _____

4. Which of the following do you use (select all that apply)?
 - Twitter
 - Facebook
 - Myspace
 - Snapchat
 - WhatsApp
 - Instagram
 - Other (please specify): _____

5. How often do you use these social networks?
 - Several times a day
 - A few times a day
 - A few times a week
 - Rarely

6. How do you access these social networks (select all that apply)?
 - From my mobile phone
 - From my tablet
 - From my computer
 - From a University desktop computer

7. Please select which of the following do you think is NOT Cyberbullying (select all that apply):
 - Calling someone rude names online.
 - Sharing an embarrassing story or pic about someone online.
 - Sharing an embarrassing video of someone online.
 - Creating a website to embarrass someone.
 - Recording a fight you witnessed and sharing it with your friends.
 - Calling someone a silly nickname that they don't mind.
 - Telling other people someone's secret online.

- Sending rude messages or pics to someone.
- Using someone's mobile phone to get them into trouble.
- Pretending to be someone online so you can talk to their friends.
- Pretending to be someone online and saying things to get them into trouble.
- Adding a rude comment on someone's picture or post

8. Have you ever been cyberbullied?

- Yes
- No (skip to question 11)
- Not sure

9. When was this?

- Within the last 3 months
- Within the last 6 months
- Within the last year
- Over a year ago

10. How did this cyberbullying occur? (select all that apply)

- They sent offensive text messages to me
- They said something offensive about me online
- They shared an offensive picture about me online
- Other (please explain): _____

11. What did you do (select all that apply)?

- I told my parents
- I told a friend
- I just ignored them
- I told them to stop
- I retaliated (please explain): _____
- I didn't know what to do
- I didn't do anything because I was too scared/upset

12. Has this cyberbullying stopped?

- Yes
- No

13. Have you witnessed a friend being cyberbullied?

- Yes
- No (skip to question 14)

14. How did this cyberbullying occur? (Select all that apply)

- They sent offensive text messages to my friend
- They posted an offensive tweet about my friend
- They sent my friend offensive private Facebook messages
- They posted something rude on my friend's Facebook wall

- They said something rude about my friend on their own wall
- They sent offensive messages to my friend on Snapchat
- They sent offensive messages to my friend on WhatsApp
- They shared rude pictures about my friend on Instagram
- Other (please explain): _____

15. What did you do (select all that apply)?

- I reported it to someone in authority (e.g. a teacher, school's administration, etc.)
- I told my parents
- I told another friend
- I told my friend to ignore them
- I told the bullies to stop
- I told my friend to retaliate
- I retaliated on behalf of my friend
- I didn't know what to do
- I didn't do anything because I was scared they would bully me too

16. Have you ever cyberbullied someone?

- Yes
- No (skip to question 16)

17. How did this cyberbullying occur? (Select all that apply)

- I sent offensive text messages to them
- I posted an offensive tweet about them
- I sent them offensive private Facebook messages
- I posted something rude on their Facebook wall
- I said something offensive about them on my Facebook wall
- I sent offensive messages to them on Snapchat
- I sent offensive messages to them on WhatsApp
- I shared rude pictures about them on Instagram
- Other (please explain): _____

18. Please select all of following statements that you agree with:

- I don't know what cyberbullying is.
- Cyberbullying is no big deal.
- Friends of mine have been cyberbullied.
- We've had cyberbullying incidents in my secondary school.
- I have cyberbullied others.
- I have said nasty things to others online, but don't consider it cyberbullying.
- I have been cyberbullied by a close friend.
- I have had someone steal my password/mobile phone and pretend to be me.
- I sent a joke to someone, but they thought it was cyberbullying.

- I've cyberbullied someone with my friends just for fun
- Others have said mean things to or about me online, but I don't consider it cyberbullying.

19. Are you happy to talk about cyberbullying with the researcher?

- Yes
- No

20. If there was an app that detects when people are being cyberbullied, would you use it?

- Yes
- No

21. Finally, in about ten words tell us what you think of cyberbully?

INFORMATION FOR RESEARCH PARTICIPATION

Study Title: *Detection and Prevention of Cyberbullying in Social Networks*

Principal Investigator: Semiu Salawu

Please read this permission form carefully and ask as many questions as you like in order to help you decide whether or not to participate in this research study. Your participation is entirely voluntary and there is no penalty or consequence for choosing not to participate. You are free to ask questions at any time before, during, or after your participation in this research. Even if you decide to participate, you are free to withdraw from the study at any time without penalty or reason.

Dear Prospective Student Participant,

We would like to invite you to take part in our research study. This form has important information about the reason for doing this study, what we will ask you to do, and the way we would like to use information about you if you choose to take part in the study.

Why are we doing this study?

You are being asked to take part in a research study to help reduce cyberbullying. This research is aimed at creating a new smartphone app. This app will help people that are being bullied online and through their phone by detecting when bullying messages are being sent. We are, therefore looking for volunteers such as yourself to help us identify what they would and would not like in this app. Any first year undergraduate student can take part, and you do not need to have been bullied or know someone that has been bullied to take part.

What will I be asked to do in this study?

We will first ask you to complete the attached questionnaire. This will allow us to get to know a little bit about you. We will review all questionnaire responses to decide who we will then invite to the next phase interview in which we will discuss what you think about cyberbullying. Due to time constraints and the fact that we need to talk to a range of different students, we will not be able to interview **everyone** who completes the questionnaire but, at the end of this study, once we have developed the smartphone app, you will be given an opportunity to test it out and tell us what you think of it, regardless of whether you took part in the interview or not.

If you are asked to take part in the interview, it will last up to 60 minutes and will be held onsite at Aston University. The discussion will mainly be about online bullying and how we can help people that are being bullied online and through their phones. This discussion will be audio recorded so we do not miss any of your contributions. We will then use these recordings in our work to create the app. Your name and details will be removed when using these recordings.

After the interview, you may be invited to take part in some group activities. In these activities, you (along with other students) will take part in what is called a participatory design session which is aimed at capturing your thoughts about the proposed mobile app, its interface and features. Three participatory design sessions are currently planned, and you may be invited to take part in all sessions. Each session can last up to two hours. These sessions will also

take place onsite at Aston University and will be audio and video recorded. Just like the interview recordings, we will use the activities' recordings to help us create the app, and your name and details will be removed when using these recordings. All recordings will be kept securely at Aston University only until the end of the project after which they will all be destroyed. In addition, for the video recordings, the video camera will not be directed at you, rather it will be pointed at a paper surface representing the app's user interface.

What are the possible risks or discomforts to taking part?

Your participation in this study may involve the following risks:

- You may get tired during the tasks. The sessions are designed to include a break after every 30 minutes, and you can rest/take a break at any time if so required.
- You may feel emotional or upset when talking about cyberbullying and some aspects of it. If this is the case, you can inform our researcher at any time if you want to take a break or stop participation. Also if the researcher feels that it is not in your best interests to continue participating, you will be removed from the study and provided details of the University's Counselling and Mental Wellbeing services and other relevant support organisations. You can also choose to stop taking part in the study at any time without having to give a reason.
- If, during discussion, you reveal something that would necessitate disciplinary measures under the University's normal policies and procedures, then the researcher present will take appropriate course of action in relation to the information revealed.

What are the possible benefits for me or others?

By taking part in this study, you will be helping us gain a better understanding of cyberbullying, how it affects young people and how it can be prevented. The end result of this study will be a new app to help reduce cyberbullying. This app will first be made available to all participants and then later on to the general public via the app store. You will have contributed to creating this app that will eventually help young people all around the world. Your participation in the study will also help raise awareness in a positive way on this very important issue.

How will you protect the information you collect about me, and how will that information be shared?

All recordings will be stored securely at Aston University and will only be available to the researcher and his supervisors. All audio and video recordings will be destroyed at the end of the study, and any reports from the study will not contain any information that can be used to identify you. The results from this study may be used in publications and presentations but you will not be identifiable in any publication.

Financial Information

Participation in this study will involve no cost (beyond the donation of time) to you. You will be entitled to up to £70 worth of Amazon vouchers for taking part in this study. The vouchers will be allocated as below:

- £10 voucher for completing the interview.
- £20 voucher for completing a participatory design session.

Thus, if for example, you completed all three participatory design sessions and the interview you'll be entitled to £70 worth of Amazon vouchers. The vouchers will be given to you at the end of each session. Please note that you will only receive a voucher if you successfully complete a session.

What are my rights as a research participant?

Participation in this study is voluntary. You may stop participating in this study at any time. If you decide not to be involved in this study, this will not affect the relationship you have with the University in any way. Your grades will not be affected if you choose not to be in this study. Equally, participation in this study will not accrue any course credits.

It is important to note that, if you decide to withdraw from this study, your contribution up to that point cannot be withdrawn where it has been part of a group activity.

Who can I contact if I have questions or concerns about this research study?

If you have any questions, you may contact the researcher at:

Semiu Salawu (PhD Researcher), Room MB267, School of Engineering & Applied Science, Aston University.

e-Mail: salawusd@aston.ac.uk

Interview Script

Background

1. Hi, my name is Semiu Salawu and I'm doing research on detecting cyberbullying on social media here in Aston University. I would like to ask you some questions about your views, personal experiences and observations on cyberbullying and, crucially, your opinions on how cyberbullying can be reduced.
2. This interview will take roughly about 60 minutes but we can break or stop at any time you want.
3. My aim is to use the information you provide to me to assist me in developing a new app that can help reduce cyberbullying by first detecting when someone is being bullied online and then taking action on behalf of the victim. If you are uncomfortable answering any of my questions, please just let me know and we can move on to the next question.

Demographic Info

4. [Confirm interviewee name, age and university year]
5. What do you understand by cyberbullying?
6. Have you ever experienced cyberbullying or witnessed someone being cyberbullied?
7. Can you tell me what happened during the incident?

Views on Cyberbullying

8. If you say something bad about someone online but you don't really mean it, do you consider that cyberbullying?
9. What about if that person has been nasty to you before or started it?
10. How can you tell if someone did not like what you said about them online?
11. How important would you say preventing cyberbullying is to you? [*Very, Somewhat, Don't Know, Not so much, I don't really care about it*]
12. Tell me more on why you gave it that level of importance?
13. What do you think are the effects of cyberbullying?

Social Media Use

14. Which social media sites/apps do you use?
15. Why do you prefer those sites?
16. What do you mainly use these sites for?
17. How many times in a day do you use these sites?
18. How long have you been using these sites?
19. Have you ever seen anyone being bullied on any of these sites?
20. Have you ever been bullied on any of these sites?
21. Have you bullied anyone on any of these sites before?
22. Do you use these sites via an app on your mobile phone or on the computer?
23. If these sites were somehow blocked on your phone and computer, what will you do?
24. When you were still in secondary school, were there any restriction on the type of websites and apps that you could use on your home computer and phone?
25. If they were restrictions, who put these restrictions in place and why?
26. Did you attempt to bypass these restrictions?
27. If yes, were you successful?

Preventing Cyberbullying

28. If someone you know is being cyberbullied, would you do anything about it?
29. If yes, what would you do?
30. Do you know of any cyberbully that was made to stop by the person being bullied?
31. What did the person do?
32. If you were in the position of the bully, would these actions discourage you from continuing the bullying?
33. How do you think cyberbullying can be reduced?

Cyberbullying Prevention App

34. Have you ever reported anyone or know anyone that got reported for cyberbullying?
35. Could the reporting process be made easier and how?
36. Can you think of new features to add to the social media sites/apps you use to help reduce cyberbullying?
37. Do you know of any software or app that can help people that are being cyberbullied?
38. If you are told to design an app to help reduce cyberbullying, what kind of features would you include in the app?
39. Would you want the app to first check all received messages do not contain bullying before delivering them to you?
40. If the app detects a bullying message that has been sent to you, what should it do? [*Should the message be automatically deleted? Should it be shown to you with a warning? Should any offensive words in the message be removed/replaced before it is delivered to you?*]
41. What about the senders of such messages? Should the app automatically block them?
42. Would you like the app to show you everything first and then you can tell it what to do next time you receive a message like that?
43. Would you want the app to also check the messages you are sending for bullying content before they are sent?
44. If bullying content is detected in the messages **you** are sending? What should the app do? [*Should it prevent you from sending it? Or show you a warning? Or replace any offensive word in the message?*]
45. If the app makes a mistake and wrongly thinks a good message is an offensive one or fails to spot a bullying message, would you stop using the app?
46. What would make you stop using the app?
47. Would you like the option to view all blocked messages and flag them as bullying or not in order to help the app become more accurate?
48. How important is such a feature to you?
49. If the app is capable of learning your general use of language and behaviour on social media so that it is able to filter offensive content better, would you be comfortable enabling such a feature in the app?
50. Would you consider such a feature good or bad?
51. Why do you think so?
52. Regardless of your views about this feature, would you consider such a feature a “killer feature” of the app?
53. Is there anything else you would like to add?
54. [Thank interviewee for their time].

APPENDIX B.14: Consent Form for First-Year Undergraduates

CONSENT FORM

for

participation in the research study entitled:

Detection and Prevention of Cyberbullying in Social Networks

Name of Researcher:	Semiu Salawu
Participant ID:	

Please initial the boxes if you agree (DO NOT tick). One signed copy of the form will be for you and the other will be kept by the researcher.

1. I have read and understood the information sheet dated for the above project. I know that I can ask the researcher questions about the project.	<input type="checkbox"/>
2. I know that taking part is voluntary and I can stop taking part in the project at any time without giving reasons.	<input type="checkbox"/>
3. I understand that all information I give to the project will be kept private and my name and details will not be publicised.	<input type="checkbox"/>
4. I understand that I will be asked to complete a questionnaire to find out more about me and how I can help the project.	<input type="checkbox"/>
5. I understand that, depending on the results of the questionnaire, I may or may not be invited to take part in other parts of the project.	<input type="checkbox"/>
6. I agree to complete the pre-study questionnaire.	<input type="checkbox"/>

Student's Name (printed) and Signature

Date

Name of Researcher Obtaining Consent and Signature

Date

CONSENT FORM

for

participation in the research study entitled:

Detection and Prevention of Cyberbullying in Social Networks

Name of Researcher:	Semiu Salawu
Participant ID:	

Please initial the boxes if you agree (DO NOT tick). One signed copy of the form will be for you and the other will be kept by the researcher.

1. I have read and understood the information sheet for the above project. I know that I can ask the researcher questions about the project.	<input type="checkbox"/>
2. I know that taking part is voluntary and I can stop taking part in the project at any time without giving reasons.	<input type="checkbox"/>
3. I understand that all information I give to the project will be kept private and my name and details will not be publicised.	<input type="checkbox"/>
4. I understand that my participation will be audio and/or recorded.	<input type="checkbox"/>
5. I understand that, if I disclose information that would require disciplinary action under Aston University's normal policies and procedures, appropriate action will be taken by the University.	<input type="checkbox"/>
6. I agree for anonymised extracts from the audio and/or video recording to be used in any reports, publications or events where results from the study will be used.	<input type="checkbox"/>
7. I agree to take part in the above study.	<input type="checkbox"/>

Student's Name (printed) and Signature

Date

Name of Researcher Obtaining Consent and Signature

Date

APPENDIX B.15: Coding Tables for Interviews' Emergent Themes

Theme	Cyberbullying occurrence intensifies in early teenhood and extends into the late teens.
Sample Quotes	Coded As
<p><i>"The effect that one person's comment can have on you, it can be huge really. A lot of people could end up killing themselves regarding it".</i></p> <p><i>"I think it's in some ways; it's worse than school bullying".</i></p> <p><i>"It's something that you can't switch off. Bullying face to face [...] It's something that you can escape but with cyberbullying its really bad [...]"</i></p> <p><i>"I feel like cyberbullying is a subject that isn't given much awareness about and especially people who are affected by it, are affected very deeply".</i></p>	<ol style="list-style-type: none"> 1. Worse than face to face 2. Pervasive 3. Not given adequate attention
<p><i>"The mean, malicious stuff didn't continue past year 12".</i></p> <p><i>"I think there was a big age thing about it. I think a lot of it was start of secondary schools, I'm thinking year seven, year eight or toward year 10, 11".</i></p>	<ol style="list-style-type: none"> 1. Cyberbullying is most frequent in years 7-10 4. Drops off in sixth form

Theme	Appearance and identity are common bullying themes.
Sample Quotes	Coded As
<p><i>"They were posting comments about her weight or bad hair or things like that".</i></p> <p><i>"It was Facebook. They were just calling her names, and there was a picture of her. She doesn't look really great in it. People thought she looked ugly basically and so they made fun of it".</i></p> <p><i>"Do you know Breaking Bad [TV series]. Walter White's son, Junior has Cerebral Palsy in the series, and I have that, so they use to call me Junior. Funny thing is we are all now friends".</i></p> <p><i>"I said I was a mixed-race person that's when the abuse came in [...]"</i></p>	<ol style="list-style-type: none"> 1. Physical appearance 2. Accent 3. Disability 4. Racial abuse

Theme	Cyberbullying on Facebook and Twitter is more public compared to Snapchat and Instagram, where it is more personal and targeted.
Sample Quotes	Coded As
<p><i>"Twitter is what I'd go on mostly. It's quite topical It's what's happening right now".</i></p> <p><i>"Probably, Snapchat the most. I think is just really accessible quite easily. It's short 10 seconds, even less than that".</i></p> <p><i>"I use Facebook Messenger the most, but in terms of social media things, probably Twitter".</i></p> <p><i>"I do have Facebook, but I don't use it [...] I follow some people on Twitter just to know what's going on".</i></p> <p><i>"Facebook is just for family and parents. No one I know uses Facebook for anything serious".</i></p> <p><i>"Snapchat is a bad one I think for cyberbullying because it's very personal and once that snap's gone, it's gone".</i></p> <p><i>"She was basically stalking her on Instagram and Twitter".</i></p> <p><i>"Twitter is quite vile. It is not a nice site at all. Twitter people are just not nice. I have just like posted inspirational</i></p>	<ol style="list-style-type: none"> 1. Facebook is less popular 2. Twitter - topical 3. Favourite platforms - Snapchat and Instagram 4. Public bullying on Facebook 5. Snapchat and Twitter – offensive messages sent directly (DM)

Theme	Cyberbullies are often known to their victims.
Sample Quotes	Coded As
<p><i>"I was probably a really nasty person to, I guess, one of my ex-girlfriends [...] she was utterly bombarded by me like five or six times a day, [...] it really wasn't until she screamed at me to leave her alone that I realised what an a*s I have been".</i></p> <p><i>"I sent loads of emails. God. Like, say I was a random person. That I was going to go to his house. I was going to blow him up. Really bad. Awful, awful, awful stuff".</i></p> <p><i>"[...] they were in my class. it was awful, what they did say. I did know it was them, there wasn't anonymity involved".</i></p> <p><i>"I've been slightly cyberbullied, but it was by an ex-boyfriend. It was on a couple of different platforms. It was Facebook Messenger and Tumblr".</i></p>	<ol style="list-style-type: none"> 1. Bullied by ex 2. Bullying an ex 3. Known to victims

Theme	Fear of reprisals and inadequate responses discourages cyberbullying reporting.
Sample Quotes	Coded As
<p><i>"I didn't want to go to school because I know there'd be a chance they'd tell my parents, and also there's definitely an element of not wanting to be a grass. I typically tell my friends. Not that they could do much about it because they were also typically being cyberbullied at the same time".</i></p> <p><i>"I typically do report things that I think have been particularly harmful. If it looks like a lot of stuff could be focused at me if I intervened then I will report anonymously. I just don't want to have that sort of stuff directed towards me".</i></p> <p><i>"We never did really talked to the teachers. You go to the teachers, what are they going to do? As soon as they tell these people off, it's just going to start right back up again and maybe worse because now they actively hate you at the same time".</i></p> <p><i>"Teachers don't know how to deal with it. The teachers aren't educated properly about what it's like to grow up in the time like us. Some of them want to help; some of them don't care".</i></p>	<ol style="list-style-type: none"> 1. Anonymous reporting 2. Fear of reprisals (can lead to physical bullying) 3. Stigma for reporting <p>Teachers are ill-equipped</p>

Theme	School should intensify cyberbullying prevention efforts.
Sample Quotes	Coded As
<p><i>"We did have a few assemblies about cyberbullying [...]".</i></p> <p><i>"[...] There were no posters or anything. Nothing. Not that I remember. If there was, it didn't stand out enough for me to see it".</i></p> <p><i>"There were blockers on the Websites and stuff. There were no social media, so no Facebook on the WiFi".</i></p>	School's effort at fighting cyberbullying

Theme	Relevant advice and punitive actions are the critical features for the proposed app.
Sample Quotes	Coded As
<p><i>"I think if advice were to be given and used on the app, it would be important whoever is writing the advice. Wherever it's coming from, they understand not just the victim's mentality but the bully's mentality as well [...]"</i></p> <p><i>"I wouldn't like it to be too childish [...] If I see anything childish, then I'll just ignore it. [...]"</i></p>	<ol style="list-style-type: none"> 1. Relevant advice 2. Pitched at the right level
<p><i>"[...] if they're given an official warning that they've been blocked from using the app, not by that person, but in general by like Facebook. Yes, then they would be like - Oh okay, I'm going to be careful right now".</i></p>	<ol style="list-style-type: none"> 1. Ban bullies 2. Report bullies

Theme	Young people would rather report cyberbullying anonymously than get directly involved
Sample Quotes	Coded As
<p><i>"Is this going to come back and hurt me in a certain way? If it would be more like yes. I wouldn't. I'm not an online superhero. That's not me".</i></p> <p><i>"I'm not sure I would be honest. I like to think I would but the bystander factor, meaning people generally just tend to ignore things that aren't happening to them or aren't important to them. I like to think I would, but I'm not sure".</i></p> <p><i>"If there was a way to report it that was anonymous or wouldn't get me involved, I would probably report it".</i></p>	<ol style="list-style-type: none"> 1. Report anonymously 2. Avoiding confrontation 3. Staying out of it

APPENDIX C.1: Links to YouTube videos on Participatory Design provided to Co-Designers.

https://www.youtube.com/watch?v=DF2sZ_EC4PU

https://www.youtube.com/watch?v=2DV_LHq_pPs

<https://www.youtube.com/watch?v=U3Hn-sONiRg>

<https://www.youtube.com/watch?v=k7CX2JYxfE8>

End User Licence Agreement

Please read this EULA carefully, as it sets out the basis upon which we license the App for use.

By clicking "Agree" below, you agree to be bound by the provisions of this EULA. If you do not agree to be bound by the provisions of this EULA, you must click the "Disagree" button below and promptly uninstall the App.

App store terms also apply

The ways in which you can use the App may also be controlled by the rules and policies of the relevant app store from which you have downloaded the App.

1. We grant you license to install and use this App on compatible devices.
2. The App is protected by the copyright laws of the United Kingdom and other countries, and we retain all intellectual property rights in the App. You may not separately publish, sell, market, distribute, lend, lease, rent, or sublicense the App code including the license key. However, this license is not to be construed as prohibiting or limiting any fair use sanctioned by copyright law, such as permitted library and classroom usage.

Limited Warranty

3. We warrant that the App will provide the features and functions generally described in the product specification on the App store where you installed it from and in the product documentation.
4. We have taken reasonable steps to keep the App free of viruses, spyware, "back door" entrances, or any other harmful code. The App will not download or install patches, upgrades, or any third party App without getting your permission. We will not intentionally deprive you of your ability to use any features of the App or access to your data.
5. You will be required to create an account to use this app. The account created will be securely stored.
6. This App may collect minimal information about how you use the App; this information will be securely stored and only used to provide the functionalities of the App. At no point will this information be passed on to a third party for the intention of advertising or marketing
7. In order to function fully, you will be required to authorize this App to access your social accounts. This authorization is done via your social account provider. This app will not see or store your social account passwords. Once authorized, the App will be able to access the information in your social account. Crucially, the App can **delete** messages, posts, status updates, etc. from your social account. Anything deleted can be viewed in Deleted Messages but cannot be re-instated to your social account. You can de-authorize this app from accessing your social accounts at any time.

8. We do not warrant that the App or your ability to use it will be uninterrupted or error-free. To the extent permitted by applicable law, we disclaim any implied warranty of merchantability or fitness for a particular purpose.

Limitation on Liability

9. The App is created as part of educational work and as such is provided on an "AS IS" basis and to the maximum extent permitted by applicable law, this material is provided AS IS AND WITH ALL FAULTS. To the fullest extent permitted by applicable law, we disclaim all liabilities, warranties and conditions, either express, implied or statutory, including, but not limited to, any (if any) implied warranties, duties or conditions of merchantability, of fitness for a particular purpose, of accuracy or completeness of responses, of results, of workmanlike effort, of lack of viruses, and of lack of negligence. In no event will any author, developer, licensor, or distributor of this App be liable to any other party for the cost of procuring substitute goods or services, lost profits, loss of use, loss of data, or any incidental, consequential, direct, indirect, punitive, or special damages whether under contract, tort, warranty, or otherwise, arising in any way out of this or any other agreement relating to this material, whether or not such party had advance notice of the possibility of such damages.

General Provisions

10. If any part of this agreement is found to be invalid or unenforceable, the remaining terms will stay in effect. This agreement does not prejudice the statutory rights of any party dealing as a consumer.

11. This agreement does not supersede any express warranties we made to you. Any modification to this agreement must be agreed to in writing by both parties.

12. This agreement will apply from the date of the installation of the App.

Privacy Policy

This App (BullStop) is the output of PhD research at Aston University, Birmingham. It is provided as a free App. The App is designed to help young people combat cyberbullying in a proactive manner. This page is to inform users regarding the policies governing the collection, use, and disclosure of Personal Information when using the App. If you choose to use the App, then you agree to the collection and use of information in relation to this policy. The Personal Information that is collected is used for providing, evaluating and improving the App. Your information will not be used or shared with anyone except as described in this Privacy Policy.

Information Collection and Use

To use the App, you will be required to create an account by providing an email address and password. The information requested will be securely stored and is used to provide you with personalisation within the App.

In order to fully function, you will be required to authorise this App to access your social media accounts. This authorisation is done via your social media account provider. The App will not see or store your social media account passwords. Once authorised, the App will be able to access information in your social media account. Crucially, the App can **delete** messages, posts, status updates, etc. from your social media account. Anything deleted can be viewed in Deleted Messages but cannot be re-instated to your social media account. You can de-authorise this App from accessing your social media accounts at any time.

After a period of at least 10 days of using the App, you may be asked to complete an online questionnaire about the App. Your questionnaire responses are not stored against any personal identifiable information. Rather a unique identifier is generated and your questionnaire responses are stored against this identifier.

Service Providers

The App may employ third-party companies and individuals due to the following reasons:

- to facilitate the services provided by the App;
- to provide the services on our behalf;
- to perform App-related functions; or
- to assist in analysing how the App is used.

You should be aware that some of these third parties may have access to the same data available to the App through your usage of the app. This is to allow them to perform the tasks assigned to them on the App's behalf. They are, however, obligated not to disclose or use the information for any other purpose.

Usage and Log Data

The App collects information about how you use the App, such as screen navigation, button taps, and swipes. Other information that may be collected include your device Internet Protocol ("IP") address, device name, operating system version, the configuration of the app, the time and date of your use of the App. Any textual data that you input into the app is not monitored. None of the data

collected can be used to identify you: rather a unique randomly generated identifier is associated with the collected data. All collected information will be securely stored and only used to evaluate and improve the App. At no point will this information be passed on to a third party for the intention of advertising or marketing. A third-party service provider (Firebase Analytics) is used to collect and analyse this data. The Firebase Analytics privacy policy is available [here](#).

Security

We value your trust in providing us with your Personal Information, thus we use commercially acceptable means of protecting it. We have put secured processes in place to ensure your data is adequately protected and not liable to data breaches.

External Links

The App may contain links to other sites. If you click on a third-party link, you will be directed to that site. Note that these external sites are operated by third parties and we have no control over and assume no responsibility for the content, privacy policies, or practices of any third-party sites or services.

Changes to This Privacy Policy

We may update our Privacy Policy from time to time. Thus, you are advised to review this page periodically for any changes. We will notify you of any changes to this Privacy Policy via email. Any changes are effective immediately after they are posted on this page.

Contact Us

If you have any questions or suggestions about this Privacy Policy, do not hesitate to contact the developer at salawusd@aston.ac.uk.

APPENDIX E.1: Instructions for Lab-Based Evaluation Study

BullStop Hands-On: Tasks to Perform (Parent-Child)

The app currently syncs with Twitter every 5 minutes so you should allow that much time for new messages to appear in the app.

1. You will be provided with an android phone/tablet with BullStop already installed. Launch the BullStop app. You will be prompted to create a BullStop account. Use the details below:

Parent:

Email: bullstoptest1@gmail.com

Password: bullstop

Child:

Email: bullstoptest2@gmail.com

Password: bullstop

2. Next, you should connect BullStop to the following Twitter account.

Parent:

Email: bullstoptest1@gmail.com

Password: bullstop

Child:

Email: bullstoptest2@gmail.com

Password: bullstop

3. BullStop starts off in Disable mode, so your first task is to configure BullStop to the protect your text messages.
4. Go to the **Messages Received** screen, there will be some messages in the inbox, open the messages and see what checkboxes BullStop assigns to the message, do you agree with the ticked checkboxes. Please, be aware that BullStop reviews messages in real time so there will be a delay of a some seconds while it analyses each opened message.
5. Go to the settings screen, set the **Detection Sensitivity** slider to any value you like. This will determine how aggressively BullStop deletes offensive messages sent to you. How does the Detection Sensitivity setting affect the way these messages are treated? Were any messages moved to Deleted Messages. Change the Detection Sensitivity slider value to a high value (maximum is 20), how has this affected the received messages?
6. Go to the **Manage Contacts** screen. Find the contact "**bullstoptest3**" and block it. Inform the researcher once you have done this and he will send you some test messages as **bullstoptest3**.
7. Go to the Received Messages screen. Confirm that no new messages have been received from "**bullstoptest3**".

APPENDIX E.1: Instructions for Lab-Based Evaluation Study

8. Open the **Twitter** app on the mobile device. The app is already configured with the account detailed above. Send some direct messages (DM) to the other participant.
9. Return to BullStop and go to the **Sent Messages** screen. Confirm that you can see the messages you sent to the other participant in the list. Open the messages and let BullStop review the messages, do you agree with the ticked checkboxes?
10. Go to the **Deleted Messages** screen. Open the deleted messages (if any) and let BullStop review the messages, do you agree with the ticked checkboxes?
11. Go to the **Message Checker** screen. Compose a message and check the message for offensive content. Do you agree with the ticked checkboxes? Send a message via the Message Checker to the other participant?
12. Go to the **Reports** screen. Verify the correctness of the reported figures.
13. Go to the **Settings** screen. Review the various settings available.
14. Go through the **Tour** screens. Can you think of anything that would make the tour better?
15. Explore the app as you wish until we are ready to discuss it as a group.

APPENDIX E.1: Instructions for Lab-Based Evaluation Study

BullStop Hands-On: Tasks to Perform (Others)

The app currently syncs with Twitter every 5 minutes so you should allow that much time for new messages to appear in the app.

1. You will be provided with an android phone/tablet with BullStop already installed. Launch the BullStop app. You will be prompted to create a BullStop account. Use the details below:

Email: bullstoptest1@gmail.com

Password: bullstop

2. Next, you should connect BullStop to the following Twitter account.

Email: bullstoptest1@gmail.com

Password: bullstop

3. BullStop starts off in Disable mode, so your first task is to configure BullStop to the protect your text messages.
4. Go to the **Messages Received** screen, there will be some messages in the inbox, open the messages and see what checkboxes BullStop assigns to the message, do you agree with the ticked checkboxes. Please, be aware that BullStop reviews messages in real time so there will be a delay of a some seconds while it analyses each opened message.
5. Go to the settings screen, set the **Detection Sensitivity** slider to any value you like. This will determine how aggressively BullStop deletes offensive messages sent to you. How does the Detection Sensitivity setting affect the way these messages are treated? Were any messages moved to Deleted Messages. Change the Detection Sensitivity slider value to a high value (maximum is 20), how has this affected the received messages?
6. Go to the **Manage Contacts** screen. Find the contact "**bullstoptest3**" and block it. Inform the researcher once you have done this and he will send you some test messages as **bullstoptest3**.
7. Go to the Received Messages screen. Confirm that no new messages have been received from "**bullstoptest3**".
8. Open the **Twitter** app on the mobile device. The app is already configured with the account detailed above. Send some direct messages (DM) to the other participant.
9. Return to BullStop and go to the **Sent Messages** screen. Confirm that you can see the messages you sent to the other participant in the list. Open the messages and let BullStop review the messages, do you agree with the ticked checkboxes?
10. Go to the **Deleted Messages** screen. Open the deleted messages (if any) and let BullStop review the messages, do you agree with the ticked checkboxes?

APPENDIX E.1: Instructions for Lab-Based Evaluation Study

11. Go to the **Message Checker** screen. Compose a message and check the message for offensive content. Do you agree with the ticked checkboxes? Send a message via the Message Checker to the other participant?
12. Go to the **Reports** screen. Verify the correctness of the reported figures.
13. Go to the **Settings** screen. Review the various settings available.
14. Go through the **Tour** screens. Can you think of anything that would make the tour better?
15. Explore the app as you wish until we are ready to discuss it as a group.

APPENDIX E.2: Questions Guide for 'Lab'-based Evaluations

1. What are your initial impressions of the app?
2. How easy was it for you to use the app?
3. Did you find the app icons intuitive?
4. Were there app icons that you did not like?
5. (If yes) Which ones and how would you improve them?
6. What do you think of the app's name?
7. What do you think of the app's logo?
8. What do you think of the colours used in the app?
9. Did you take the app tour?
10. (If yes) Did you find the app tour useful?
11. What do you think about the app's accuracy in detecting offensive and bullying messages?
12. Did you correct any of the Deleted Messages?
13. (If yes) How many?
14. Did you correct any of the Received Messages?
15. (If yes) How many?
16. Did you correct any of the Sent Messages?
17. (If yes) How many?
18. What do you think of the Message Checker?
19. What do you think of the Reports?
20. What was your favourite thing in the app?
21. What was your least favourite thing in the app?
22. How would you improve the app?
23. Do you think the app will be useful to other young people?
24. If you were being cyberbullied would you use the app?
25. If yes) Why?
26. Would you use the app even if you are not being cyberbullied?
27. (If yes) Why?

Date: 16 October 2019

Dr Jo Lumsden

Student: Semiu Salawu

Dear Jo,

Study title:	Detection and Prevention of Cyberbullying on Online Social Networks
REC REF:	# 1544

Confirmation of Ethical Opinion

On behalf of the Committee, I am pleased to confirm a favourable opinion for the above research based on the basis described in the application form, protocol and supporting documentation listed below.

Approved documents

The final list of documents reviewed and approved by the Committee is as follows.

<i>Document</i>	<i>Version</i>	<i>Date</i>
Participant Information Sheet: Teenager	1.C	03/09/2019
Consent Form: Teenager: Teenager	1.C	03/09/2019
Participant Information Sheet: Parent	1.P	03/09/2019
Consent Form: Parent	1.P	03/09/2019
Participant Information Sheet: Professional	1.PR	03/09/2019
Consent Form: Professional	1.PR	03/09/2019
Recruitment Text: Aston	1	16/10/2019
Recruitment Text: Parent	1	16/10/2019
Recruitment Text: Professional	1	16/10/2019
Phase 1 App Tasks	1	16/10/2019
Phase 1 Focus Group Protocols	1	16/10/2019
Phase 1 Focus Group Questions	1	16/10/2019
Phase 1 Interview Questions	1	16/10/2019
Phase 1 Paired exploration protocols	1	16/10/2019
Phase 1 Pared exploration questions	1	16/10/2019

After starting your research please notify the University Research Ethics Committee of any of the following:

- Amendments. Any amendment should be sent as a Word document, with the amendment highlighted or showing tracked changes. The amendment request must be accompanied by a covering letter along with all amended documents, e.g. protocols, participant information sheets, consent forms etc. Please include a version number and amended date to the file name of any amended documentation (e.g. "Ethics Application #100 Protocol v2 amended 17/02/19.doc").
- Unforeseen or adverse events e.g. disclosure of personal data, harm to participants.
- New Investigators
- End of the study

Please email all notifications or queries to research_governance@aston.ac.uk and quote your UREC reference number with all correspondence.

Wishing you every success with your research.

Yours sincerely

Ali Alshukry
Acting Chair, University Research Ethics Committee

APPENDIX E.4: Invitation Email for Lab-Based Evaluation Study (Professionals).

Invitation to Participate

Evaluation of BullStop: A Mobile App for The Detection and Prevention of Cyberbullying on Social Media

Hi [name],

As we discussed, I would like to invite you to participate in a research study to evaluate a mobile app designed to help young people combat cyberbullying on social media sites.

I am looking for [*educators/cybercrime enforcement officers/child mental health professionals*] such as yourself to take part in the study. The session will last between 60 – 90 minutes and will involve you using an Android phone pre-installed with the app for a period of up to 30 minutes. After this we will then spend between 30 minutes to an hour talking about your impressions of the app and any feedback you may have. We can do the session at a local café or any venue convenient to you.

I have attached a participant information sheet (which provides additional information about the study) to this email. Please read it and feel free to ask as many questions as you need in order to decide if you wish to take part.

I have also included a consent form for you to preview. If you agree to take part, you will be required to sign a hardcopy of the consent form before we start the session.

If you like to take part, please let me know. We can then arrange a suitable date, time and location for the session.

If you have any question, please don't hesitate to contact me. Thank you.

APPENDIX E.5: Invitation Email for Lab-Based Evaluation Study (Parents).

Invitation to Participate

Evaluation of BullStop: A Mobile App for The Detection and Prevention of Cyberbullying on Social Media

Hi [name],

As we discussed, I would like to invite you to participate in a research study to evaluate a mobile app designed to help young people combat cyberbullying on social media sites.

I am looking for young people aged 11 – 17 years and their parents to take part together in the study. The session will last between 60 – 90 minutes and will involve you and your child collaboratively using an Android phone pre-installed with the app for a period of up to 30 minutes. After this we will, as a group, then spend between 30 minutes to an hour talking about your impressions of the app and any feedback you may have. We can do the session either at your home, a local café or any venue convenient to you. It is important that both you and [Child Name] *want* to take part: please don't place [Child Name] under any pressure to participate. You will be with [Child Name] at all times during the study.

I have attached two participant information sheets (these provide additional information about the study) to this email – one for you and one for [Child Name]. Please read it and feel free to ask as many questions as you or [Child Name] need in order to decide if you each wish to take part. Please discuss participation carefully with [Child Name] before deciding whether to participate.

I have included 3 consent forms for you to preview as specified below:

- Adult consent form for you;
- Parent consent form for you to complete and sign on behalf of your child (required for children under 13 years); and
- A teenager consent form (required for children 13 and over to complete by themselves).

If you and [Child Name] agree to take part, you will be asked to sign hardcopies of the appropriate consent forms before we start the session.

If you and [Child Name] would like to take part, please let me know. We can then arrange a suitable date, time and location for the session.

If you have any question, please don't hesitate to contact me. Thank you.

APPENDIX E.6: Participant Information Sheet for Lab-Based Evaluation Study (Professionals).



EVALUATION OF BULLSTOP: A MOBILE APP FOR THE DETECTION AND PREVENTION OF CYBERBULLYING ON SOCIAL MEDIA

Participant Information Sheet

Invitation

We would like to invite you to take part in a research study.

Before you decide if you would like to participate, take time to read the following information carefully and, if you wish, discuss it with others such as your family, friends or colleagues.

Please ask a member of the research team, whose contact details can be found at the end of this information sheet, if there is anything that is not clear or if you would like more information before you make your decision.

What is the purpose of this study?

You are being asked to participate in a study to evaluate the usability and usefulness of a mobile app designed to detect and combat cyberbullying on social media. The app uses novel computing technology to detect various forms of cyberbullying and take appropriate actions (such as deleting offensive messages and blocking cyberbullies and trolls) on behalf of the user. The target audience for this mobile application will be young people in the UK aged 11 – 17 years. We hope that this app will help protect vulnerable people, especially children, from the damage caused by cyberbullying.

Why have I been chosen?

You are being invited to take part in this study because as a [*profession*], we believe your input will be very valuable in assessing the mobile app from the perspective of responsible adults in the lives of the young people targeted by the app.

What will happen to me if I take part?

You will be asked to evaluate the app by using an Android smartphone to complete a series of tasks within the app. After using the app to complete the tasks, you will be asked to discuss your opinions of the app with the researcher in an interview. This interview will be audio recorded so that we can maintain an accurate record of what you thought about the app.

To protect your identity, you will be provided with a dummy account to login and use the app, and you will also be provided with a dummy Twitter account to which to connect the app. In other words, none of your personal contact information or your social media activity will be visible to the researchers or recorded as part of this study.

As the app is designed to be intuitive and simple to use, no special skills are required to use the app. It is designed for the same level of app competence as any social media app. You can however ask questions or request assistance from the researcher if required at any time. In addition, the app includes a tutorial on how to use the app which you can refer to whenever you wish.

The app is designed to record and track user actions such as screen navigation and taps. Keystrokes are however NOT recorded and so anything you type into the app is not recorded.

APPENDIX E.6: Participant Information Sheet for Lab-Based Evaluation Study (Professionals).

The entire session is expected to last between 60 – 90 minutes.

Do I have to take part?

No. It is up to you to decide whether or not you wish to take part. If you do decide to participate, you will be asked to sign and date a consent form. You will still be free to withdraw from the study at any time without giving a reason.

Will my taking part in this study be kept confidential?

Yes. A code will be attached to all the data you provide to maintain confidentiality.

Your personal data (name and contact details) will only be used if the researchers need to contact you to arrange study visits or collect data by phone. Analysis of your data will be undertaken using coded data.

The data we collect will be stored in a secure document store (paper records) or electronically on a secure encrypted mobile device, password protected computer server or secure cloud storage device.

To ensure the quality of the research, Aston University may need to access your data to check that the data has been recorded accurately. If this is required, your personal data will be treated as confidential by the individuals accessing your data.

How will the conversations that take place during the interview be recorded and the information I provide managed?

With your permission we will audio record the interview and take notes.

The recording will be typed into a document (transcribed) by the researcher. This process will involve removing any information which could be used to identify individuals e.g., names, locations, etc.

Audio recordings will be destroyed as soon as the transcripts have been checked for accuracy. We will ensure that anything you have told us that is included in the reporting of the study is anonymous.

You of course are free not to answer any questions that are asked without giving a reason.

What are the possible benefits of taking part?

While there are no direct benefits to you of taking part in this study, the data gained will contribute to knowledge that will allow us to make this app as good as it can be and, in future, design other applications to mitigate and prevent cyberbullying.

What are the possible risks and burdens of taking part?

There are minimal risks associated with participating in this study beyond that of normal everyday usage of social media. The app (BullStop) is targeting a sensitive issue and, in order to allow participants to properly evaluate its use, the app has been pre-populated with 'fake' bullying content. A small portion of these offensive messages may contain one or more of the following profane words – damn, hell, sh*t, fu*k. None of the messages include racist, homophobic, transphobic, sexist and offensive content about age, weight, physical appearance and religion.

Whilst the fake content will be carefully selected to avoid being unduly offensive, it is recognised that it could be offensive for some and/or could trigger memories of personal bullying for some participants. If any of the bullying messages make you feel uncomfortable then you should let the researcher know immediately; similarly, if the researcher observes you exhibiting behaviour indicative

APPENDIX E.6: Participant Information Sheet for Lab-Based Evaluation Study (Professionals).

of psychological discomfort, he will suspend your involvement until convinced that you are able to and wish to continue. If participation in the study causes distress, you will be directed to a range of help resources (links to which are also embedded in the app itself).

You can at any point stop your participation in the study. This does not affect any payments to which you are entitled.

What will happen to the results of the study?

The results of this study may be published/presented in scientific journals and/or presented at conferences. If the results of the study are published, your identity will remain confidential.

A lay summary of the results of the study will be available for participants when the study has been completed and the researchers will ask if you would like to receive a copy.

The results of the study will also be used in the PhD thesis of Semiu Salawu.

Expenses and payments

You will be entitled to £10 worth of Amazon vouchers for taking part in this study. No expenses payments will be made.

Who is funding the research?

This research is self-funded by the researcher (Semiu Salawu).

Who is organising this study and acting as data controller for the study?

Aston University is organising the study and acting as data controller for the study. You can find out more about how we use your information in Appendix A.

Who has reviewed the study?

This study was given a favourable ethical opinion by Aston University Research Ethics Committee.

What if I have a concern about my participation in the study?

If you have any concerns about your participation in this study, please speak to the research team and they will do their best to answer your questions. Contact details can be found at the end of this information sheet.

If the research team are unable to address your concerns or you wish to make a complaint about how the study is being conducted, you should contact the Aston University Research Integrity Office at research_governance@aston.ac.uk or telephone 0121 204 3000.

Research Team

If you have any questions, you may contact the PhD student researcher or his supervisor at:

Semiu Salawu (PhD Student), School of Engineering & Applied Science, Aston University, e-Mail: salawusd@aston.ac.uk.

Dr Jo Lumsden (Supervisor), School of Engineering & Applied Science, Aston University, e-Mail: j.lumsden@aston.ac.uk. Tel: 0121 204 3470

Thank you for taking time to read this information sheet. If you have any questions regarding the study, please don't hesitate to ask one of the research team.



Aston University takes its obligations under data and privacy law seriously and complies with the General Data Protection Regulation (“GDPR”) and the Data Protection Act 2018 (“DPA”).

Aston University is the sponsor for this study based in the United Kingdom. We will be using information from you in order to undertake this study. Aston University will process your personal data in order to register you as a participant and to manage your participation in the study. It will process your personal data on the grounds that it is necessary for the performance of a task carried out in the public interest (GDPR Article 6(1)(e)). Aston University may process special categories of data about you which includes details about your health. Aston University will process this data on the grounds that it is necessary for statistical or research purposes (GDPR Article 9(2)(j)). Aston University will keep identifiable information about you for 6 years after the study has finished.

Your rights to access, change or move your information are limited, as we need to manage your information in specific ways in order for the research to be reliable and accurate. If you withdraw from the study, we will keep the information about you that we have already obtained. To safeguard your rights, we will use the minimum personally identifiable information possible.

You can find out more about how we use your information at www.aston.ac.uk/dataprotection or by contacting our Data Protection Officer at dp_officer@aston.ac.uk.

If you wish to raise a complaint on how we have handled your personal data, you can contact our Data Protection Officer who will investigate the matter. If you are not satisfied with our response or believe we are processing your personal data in a way that is not lawful you can complain to the Information Commissioner’s Office (ICO).

APPENDIX E.7: Participant Information Sheet for Lab-Based Evaluation Study (Parents).



EVALUATION OF BULLSTOP: A MOBILE APP FOR THE DETECTION AND PREVENTION OF CYBERBULLYING ON SOCIAL MEDIA

Participant Information Sheet

Invitation

We would like to invite you to take part in a research study.

Before you decide if you would like to participate, take time to read the following information carefully and, if you wish, discuss it with others such as your family, friends or colleagues.

Please ask a member of the research team, whose contact details can be found at the end of this information sheet, if there is anything that is not clear or if you would like more information before you make your decision.

What is the purpose of this study?

You are being asked to participate in a study to evaluate the usability and usefulness of a mobile app designed to detect and combat cyberbullying on social media. The app uses novel computing technology to detect various forms of cyberbullying and take appropriate actions (such as deleting offensive messages and blocking cyberbullies and trolls) on behalf of the user. The target audience for this mobile application will be young people in the UK aged 11 – 17 years. We hope that this app will help protect vulnerable people, especially children, from the damage caused by cyberbullying.

Why have I been chosen?

You are being invited to take part in this study because you are a parent with a child that falls within the age range of the target audience for the app. As a parent, we believe your input will be very valuable in assessing how well the mobile app meets the needs of our target audience and to represent the opinions of parents. Please note that we make no assumptions about any experience you or your child may have had with cyberbullying in the past – you are not being invited because we think you or your child may have had any experience of cyberbullying.

What will happen to me if I take part?

You will be asked to evaluate the app alongside your child. In this session, you will be asked to use an Android smartphone to interact with the app to complete a series of tasks. After using the app to complete the tasks, you and your child will be asked to discuss your opinions of the app with the researcher in a paired interview. This interview will be audio recorded so that we can maintain an accurate record of what you and your child thought about the app.

To protect your identity, you will be provided with a dummy account to login and use the app, and you will also be provided with a dummy Twitter account to which to connect the app. In other words, none of your personal contact information or your social media activity will be visible to the researchers or recorded as part of this study.

As the app is designed to be intuitive and simple to use, no special skills are required to use the app. It is designed for the same level of app competence as any social media app. You and your child can

APPENDIX E.7: Participant Information Sheet for Lab-Based Evaluation Study (Parents).

however ask questions or request assistance from the researcher if required at any time. In addition, the app includes a tutorial on how to use the app which you can refer to whenever you wish.

The app is designed to record and track user actions such as screen navigation and taps. Keystrokes are however NOT recorded and so anything you type into the app is not recorded.

The entire session is expected to last between 60 – 90 minutes.

Do I have to take part?

No. It is up to you to decide whether or not you and your child wish to take part. If you do decide to participate, you will be asked to sign and date a consent form. If your child is under 13 years, you will also be required to sign and date a consent form on their behalf. You and your child will still be free to withdraw from the study at any time without giving a reason.

Will my taking part in this study be kept confidential?

Yes. A code will be attached to all the data you provide to maintain confidentiality.

Your personal data (name and contact details) will only be used if the researchers need to contact you to arrange study visits or collect data by phone. Analysis of your data will be undertaken using coded data.

The data we collect will be stored in a secure document store (paper records) or electronically on a secure encrypted mobile device, password protected computer server or secure cloud storage device.

To ensure the quality of the research, Aston University may need to access your data to check that the data has been recorded accurately. If this is required, your personal data will be treated as confidential by the individuals accessing your data.

How will the conversations that take place during the interview be recorded and the information I provide managed?

With your permission we will audio record the interview and take notes.

The recording will be typed into a document (transcribed) by the researcher. This process will involve removing any information which could be used to identify individuals e.g., names, locations, etc.

Audio recordings will be destroyed as soon as the transcripts have been checked for accuracy.

We will ensure that anything you have told us that is included in the reporting of the study is anonymous.

You of course are free not to answer any questions that are asked without giving a reason.

What are the possible benefits of taking part?

While there are no direct benefits to you of taking part in this study, the data gained will contribute to knowledge that will allow us to make this app as good as it can be and, in future, design other applications to mitigate and prevent cyberbullying.

What are the possible risks and burdens of taking part?

There are minimal risks associated with participating in this study beyond that of normal everyday usage of social media. The app (BullStop) is targeting a sensitive issue and, in order to allow participants to properly evaluate its use, the app has been pre-populated with 'fake' bullying content.

APPENDIX E.7: Participant Information Sheet for Lab-Based Evaluation Study (Parents).

A small portion of these offensive messages may contain one or more of the following profane words – damn, hell, sh*t, fu*k. None of the messages include racist, homophobic, transphobic, sexist and offensive content about age, weight, physical appearance and religion.

Whilst the fake content will be carefully selected to avoid being unduly offensive, it is recognised that it could be offensive for some and/or could trigger memories of personal bullying for some participants. If any of the bullying messages make you or your child feel uncomfortable then you should let the researcher know immediately; similarly, if the researcher observes you or your child exhibiting behaviour indicative of psychological discomfort, he will suspend your involvement until convinced that you or child are able to and wish to continue. If participation in the study causes distress, you will be directed to a range of help resources (links to which are also embedded in the app itself).

If your child has been subjected to cyberbullying in the past, you should think carefully about allowing their participation in this study so as not to risk them reliving the experience. We would strongly advise talking this over with your child first before agreeing to their participation in the study.

Also, please be aware that you or your child can at any point stop your participation in the study. This does not affect any payments to which you are entitled.

What will happen to the results of the study?

The results of this study may be published/presented in scientific journals and/or presented at conferences. If the results of the study are published, your identity will remain confidential.

A lay summary of the results of the study will be available for participants when the study has been completed and the researchers will ask if you would like to receive a copy.

The results of the study will also be used in the PhD thesis of Semiu Salawu.

Expenses and payments

You and your child will be entitled to £10 worth of Amazon vouchers each for taking part in this study. No expenses payments will be made.

Who is funding the research?

This research is self-funded by the researcher (Semiu Salawu).

Who is organising this study and acting as data controller for the study?

Aston University is organising the study and acting as data controller for the study. You can find out more about how we use your information in Appendix A.

Who has reviewed the study?

This study was given a favourable ethical opinion by Aston University Research Ethics Committee.

What if I have a concern about my participation in the study?

If you have any concerns about your participation in this study, please speak to the research team and they will do their best to answer your questions. Contact details can be found at the end of this information sheet.

If the research team are unable to address your concerns or you wish to make a complaint about how the study is being conducted, you should contact the Aston University Research Integrity Office at research_governance@aston.ac.uk or telephone 0121 204 3000.

APPENDIX E.7: Participant Information Sheet for Lab-Based Evaluation Study (Parents).

Research Team

If you have any questions, you may contact the PhD student researcher or his supervisor at:

Semiu Salawu (PhD Student), School of Engineering & Applied Science, Aston University, e-Mail: salawusd@aston.ac.uk.

Dr Jo Lumsden (Supervisor), School of Engineering & Applied Science, Aston University, e-Mail: j.lumsden@aston.ac.uk. Tel: 0121 204 3470

Thank you for taking time to read this information sheet. If you have any questions regarding the study, please don't hesitate to ask one of the research team.



Aston University takes its obligations under data and privacy law seriously and complies with the General Data Protection Regulation (“GDPR”) and the Data Protection Act 2018 (“DPA”).

Aston University is the sponsor for this study based in the United Kingdom. We will be using information from you in order to undertake this study. Aston University will process your personal data in order to register you as a participant and to manage your participation in the study. It will process your personal data on the grounds that it is necessary for the performance of a task carried out in the public interest (GDPR Article 6(1)(e)). Aston University may process special categories of data about you which includes details about your health. Aston University will process this data on the grounds that it is necessary for statistical or research purposes (GDPR Article 9(2)(j)). Aston University will keep identifiable information about you for 6 years after the study has finished.

Your rights to access, change or move your information are limited, as we need to manage your information in specific ways in order for the research to be reliable and accurate. If you withdraw from the study, we will keep the information about you that we have already obtained. To safeguard your rights, we will use the minimum personally identifiable information possible.

You can find out more about how we use your information at www.aston.ac.uk/dataprotection or by contacting our Data Protection Officer at dp_officer@aston.ac.uk.

If you wish to raise a complaint on how we have handled your personal data, you can contact our Data Protection Officer who will investigate the matter. If you are not satisfied with our response or believe we are processing your personal data in a way that is not lawful you can complain to the Information Commissioner’s Office (ICO).

APPENDIX E.8: Participant Information Sheet for Lab-Based Evaluation Study (Young People).



EVALUATION OF BULLSTOP: A MOBILE APP FOR THE DETECTION AND PREVENTION OF CYBERBULLYING ON SOCIAL MEDIA

Participant Information Sheet

Invitation

We would like to invite you to take part in a research study.

Before you decide if you would like to take part, take time to read the following information carefully and, if you wish, discuss it with others such as your family, friends or colleagues.

Please ask a member of the research team, whose contact details can be found at the end of this information sheet, if there is anything that is not clear or if you would like more information before you make your decision.

What is the purpose of this study?

You are being asked to take part in a study to assess how usable and useful a mobile app designed to detect and combat cyberbullying on social media is. The app uses new computing technology to detect types of cyberbullying and take action to protect its user (such as deleting upsetting messages and blocking cyberbullies and trolls). The target audience for this mobile application will be young people in the UK aged 11 – 17 years – i.e., people like you. We hope that this app will help protect young people from the negative effects of cyberbullying.

Why have I been chosen?

You are being invited to take part in this study because you are someone of the age this app is designed for. We believe your input will be very useful in assessing how well the mobile app meets the needs of young people like you. Please note that we are not assuming that you have any experience of cyberbullying in the past – you are not being invited because we think you may have had any experience of cyberbullying.

What will happen to me if I take part?

You will be asked to assess the app alongside your parent. In this session, you will be asked to use an Android smartphone to interact with the app to complete a series of tasks. After using the app to complete the tasks, you and your parent will be asked to discuss your opinions of the app with the researcher in a paired interview. This interview will be audio recorded so that we can accurately record what you and your parent thought about the app.

To protect your identity, you will be provided with a dummy account to login and use the app, and you will also be provided with a dummy Twitter account to which to connect the app. In other words, none of your personal contact information or your social media activity will be visible to the researchers or recorded as part of this study.

As the app is designed to be simple to use, no special skills are required to use the app. It is designed to be usable by the same people who use any social media app. You can however ask questions or request assistance from the researcher if required at any time. In addition, the app includes a tutorial on how to use the app which you can refer to whenever you wish.

APPENDIX E.8: Participant Information Sheet for Lab-Based Evaluation Study (Young People).

The app is designed to record and track user actions such as screen navigation and taps. Keystrokes are however NOT recorded and so anything you type into the app is not recorded.

The entire session is expected to last between 60 – 90 minutes.

Do I have to take part?

No. It is up to you to decide whether or not you wish to take part. If you do decide to take part, you will be asked to sign and date a consent form (if you are under 13, your parent will sign this on your behalf). You will still be free to stop taking part in the study at any time without giving a reason.

Will my taking part in this study be kept confidential?

Yes. A code will be attached to all the data you provide to make sure you can't be identified from the data.

Your personal data (name and contact details) will only be used if the researchers need to contact you to arrange study visits or collect data by phone. Analysis of your data will be undertaken using coded data.

The data we collect will be stored in a secure document store (paper records) or electronically on a secure encrypted mobile device, password protected computer server or secure cloud storage device.

To ensure the quality of the research, Aston University may need to access your data to check that the data has been recorded accurately. If this is required, your personal data will be treated as confidential by the individuals accessing your data.

How will the conversations that take place during the interview be recorded and the information I provide managed?

With your permission we will audio record the interview and take notes.

The recording will be typed into a document (transcribed) by the researcher. This process will involve removing any information which could be used to identify individuals e.g., names, locations, etc.

Audio recordings will be destroyed as soon as the transcripts have been checked to make sure they are correct.

We will ensure that anything you have told us that is included in the reporting of the study does not include your name.

You of course are free not to answer any questions that are asked without giving a reason.

What are the possible benefits of taking part?

While you won't benefit directly from taking part in this study, the information you provide us with will contribute to knowledge that will allow us to make this app as good as it can be and, in future, design other applications to reduce and/or prevent cyberbullying.

What are the possible risks and burdens of taking part?

There are very few risks associated with taking part in this study beyond that of normal everyday use of social media. The app (BullStop) is targeting a sensitive issue and, in order to allow participants to properly evaluate its use, the app has been pre-populated with 'fake' bullying content. A small portion of these offensive messages may contain one or more of the following words – damn, hell, sh*t, fu*k.

APPENDIX E.8: Participant Information Sheet for Lab-Based Evaluation Study (Young People).

None of the messages include racist, homophobic, transphobic, sexist and offensive content about age, weight, physical appearance and religion.

Whilst the fake content will be carefully selected to avoid being too offensive, we realise it could upset some people and/or could bring up memories of personal bullying for some people. If any of the bullying messages make you feel uncomfortable then you should either let your parent/guardian or the researcher know immediately; similarly, if the researcher thinks that you look like you are uncomfortable, he will pause your involvement until he thinks that you are able to and/or wish to continue. If participation in the study causes you to become upset, the researcher will point you towards to a range of help resources (links to which are also embedded in the app itself).

You can at any point stop taking part in the study. This does not affect whether or not you will be given your £10 Amazon voucher.

What will happen to the results of the study?

The results of this study may be published/presented in scientific journals and/or presented at conferences. If the results of the study are published, your identity will remain hidden.

A summary of the results of the study will be available for you if you like when the study has been completed and the researchers will ask if you would like to receive a copy.

The results of the study will also be used in the PhD thesis of Semiu Salawu.

Expenses and payments

You will be entitled to £10 worth of Amazon vouchers for taking part in this study. No expenses payments will be made.

Who is funding the research?

This research is self-funded by the researcher (Semiu Salawu).

Who is organising this study and acting as data controller for the study?

Aston University is organising the study and acting as data controller for the study. You can find out more about how we use your information in Appendix A.

Who has reviewed the study?

This study was given a favourable ethical opinion by Aston University Research Ethics Committee.

What if I have a concern about my participation in the study?

If you have any concerns about your taking part in this study, please speak to the research team and they will do their best to answer your questions. Contact details can be found at the end of this information sheet.

If the research team are unable to address your concerns or you wish to make a complaint about how the study is being conducted, you should contact the Aston University Research Integrity Office at research_governance@aston.ac.uk or telephone 0121 204 3000.

Research Team

If you have any questions, you may contact the PhD student researcher or his supervisor at:

Semiu Salawu (PhD Student), School of Engineering & Applied Science, Aston University, e-Mail: salawusd@aston.ac.uk.

APPENDIX E.8: Participant Information Sheet for Lab-Based Evaluation Study (Young People).

Dr Jo Lumsden (Supervisor), School of Engineering & Applied Science, Aston University, e-Mail: j.lumsden@aston.ac.uk. Tel: 0121 204 3470

Thank you for taking time to read this information sheet. If you have any questions regarding the study, please don't hesitate to ask one of the research team.



Aston University takes its obligations under data and privacy law seriously and complies with the General Data Protection Regulation (“GDPR”) and the Data Protection Act 2018 (“DPA”).

Aston University is the sponsor for this study based in the United Kingdom. We will be using information from you in order to undertake this study. Aston University will process your personal data in order to register you as a participant and to manage your participation in the study. It will process your personal data on the grounds that it is necessary for the performance of a task carried out in the public interest (GDPR Article 6(1)(e)). Aston University may process special categories of data about you which includes details about your health. Aston University will process this data on the grounds that it is necessary for statistical or research purposes (GDPR Article 9(2)(j)). Aston University will keep identifiable information about you for 6 years after the study has finished.

Your rights to access, change or move your information are limited, as we need to manage your information in specific ways in order for the research to be reliable and accurate. If you withdraw from the study, we will keep the information about you that we have already obtained. To safeguard your rights, we will use the minimum personally identifiable information possible.

You can find out more about how we use your information at www.aston.ac.uk/dataprotection or by contacting our Data Protection Officer at dp_officer@aston.ac.uk.

If you wish to raise a complaint on how we have handled your personal data, you can contact our Data Protection Officer who will investigate the matter. If you are not satisfied with our response or believe we are processing your personal data in a way that is not lawful you can complain to the Information Commissioner’s Office (ICO).



EVALUATION OF BULLSTOP: A MOBILE APP FOR THE DETECTION AND PREVENTION OF CYBERBULLYING ON SOCIAL MEDIA

Consent Form

Name of Chief Investigator: Semiu Salawu

Please initial boxes

1.	I confirm that I have read and understand the Participant Information Sheet (Version #1.P 03/09/2019) for the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily.	
2.	I understand that my participation is voluntary and that I am free to withdraw at any time, without giving any reason and without my legal rights being affected.	
3.	I agree to my personal data and data relating to me collected during the study being processed as described in the Participant Information Sheet	
4.	I agree to my interview being audio recorded and to anonymised direct quotes from me being used in publications resulting from the study.	
5.	I agree to my anonymised data being used by research teams for future research.	
6.	I agree to my personal data being processed for the purposes of inviting me to participate in future research projects. I understand that I may opt out of receiving these invitations at any time.	
7.	I agree to take part in this study.	

Name of participant

Date

Signature

Name of Person receiving consent.

Date

Signature

APPENDIX E.10: Participant Information Sheet for Lab-Based Evaluation Study (Parents).



EVALUATION OF BULLSTOP: A MOBILE APP FOR THE DETECTION AND PREVENTION OF CYBERBULLYING ON SOCIAL MEDIA

Consent Form

Name of Chief Investigator: Semiu Salawu

Please initial boxes

1.	I confirm that I have read and understand the Participant Information Sheet (Version #1.X 03/09/2019) for the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily.	
2.	I understand that my participation is voluntary and that I am free to withdraw at any time, without giving any reason and without my legal rights being affected.	
3.	I agree to my personal data and data relating to me collected during the study being processed as described in the Participant Information Sheet	
4.	I understand that if during the study I or my child tell the research team something that causes them to have concerns in relation to his/her health and/or welfare they may need to breach my confidentiality.	
5.	I agree to my interview being audio recorded and to anonymised direct quotes from me being used in publications resulting from the study.	
6.	I agree to my anonymised data being used by research teams for future research.	
7.	I agree to my personal data being processed for the purposes of inviting me to participate in future research projects. I understand that I may opt out of receiving these invitations at any time.	
8.	I agree to take part in this study.	

Name of participant

Date

Signature

Name of Person receiving consent.

Date

Signature

APPENDIX E.11: Participant Information Sheet for Lab-Based Evaluation Study (Parental – on Behalf on Child).



EVALUATION OF BULLSTOP: A MOBILE APP FOR THE DETECTION AND PREVENTION OF CYBERBULLYING ON SOCIAL MEDIA

Consent Form

Name of Chief Investigator: Semiu Salawu

Please initial boxes

1.	I confirm that I have read and understand the Participant Information Sheet (Version #1.X 03/09/2019) for the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily.	
2.	I understand that my child’s participation is voluntary and that I am free to withdraw my child’s participation at any time, without giving any reason and without my or my child’s legal rights being affected.	
3.	I agree to my child’s personal data and data relating to my child collected during the study being processed as described in the Participant Information Sheet.	
4.	I understand that if during the study my child tells the research team something that causes them to have concerns in relation to his/her health and/or welfare they may need to breach my child’s confidentiality.	
5.	I agree to my child’s interview being audio recorded and to anonymised direct quotes from my child being used in publications resulting from the study.	
6.	I agree to my child’s anonymised data being used by research teams for future research.	
7.	I agree to my child’s personal data being processed for the purposes of inviting my child to participate in future research projects. I understand that my child may opt out of receiving these invitations at any time.	
8.	I agree for my child to take part in this study.	

Name of Parent/Legal Guardian Date

Signature

Name of Person receiving consent. Date

Signature

APPENDIX E.12: Coding Tables for 'Lab'-Based Evaluation Study Emergent Themes

Theme	Easy to use
Sample Quotes	Coded As
<p><i>"I like the simplicity of the UI. I don't like cluttered apps with too many things going on the screen".</i></p> <p><i>"It's simple. The whole app is simple. I like the settings".</i></p> <p><i>"I like the look. It looks very professional and clean".</i></p> <p><i>"I like the design and the colours. It's very subtle, not in your face".</i></p> <p><i>"I like the pictures of the kids on the home page. And the settings, it's straightforward."</i></p> <p><i>"It's simple. It's like Twitter but also different. I get it from the beginning. I like the settings".</i></p> <p><i>"It's easy to set up an account initially, and then to log in, it's quite easy and quick. There's not a lot of confusing stuff, and it's really quick to create an account. That's very good".</i></p> <p><i>"I like the way everything is set out".</i></p> <p><i>"I think it was quite good for a techy challenged person, easy to use, and straight forward, and not too busy".</i></p> <p><i>"Easy to set up the account and it connects well with Twitter".</i></p>	<ol style="list-style-type: none"> 1. Simple interface 2. Easy to use
<p><i>"I like the tour. I generally struggle with apps or anything techy, but with a few screens of the tour, I understood the app completely".</i></p> <p><i>"My initial impression was very good. You start the app; you can see the tour. As a user, or even as perhaps a guardian, I can go through the tour and understand very quickly what the app is trying to achieve. That is a very good start because some apps can be confusing when you first open them, and that makes this good".</i></p>	<ol style="list-style-type: none"> 1. Helpful 2. Guides the user.

Theme	Well designed
Sample Quotes	Coded As
<p><i>"I like the look. It looks very professional and clean".</i></p> <p><i>"It certainly looks like a paid app I would have downloaded from the app store".</i></p> <p><i>"It looks very professional. Like apps [that] I have paid for before".</i></p>	<ol style="list-style-type: none"> 1. High quality 2. Well designed 3. Appealing
<p><i>"I like the tour. I generally struggle with apps or anything techy, but with a few screens of the tour, I understood the app completely".</i></p>	<ol style="list-style-type: none"> 1. Helpful 2. Guides the user

APPENDIX E.12: Coding Tables for ‘Lab’-Based Evaluation Study Emergent Themes

<i>“My initial impression was very good. You start the app; you can see the tour. As a user, or even as perhaps a guardian, I can go through the tour and understand very quickly what the app is trying to achieve. That is a very good start because some apps can be confusing when you first open them, and that makes this good”.</i>	
--	--

Theme	Appropriate branding
Sample Quotes	Coded As
<p><i>“I like the shield. It has a strong message about what it’s stopping”.</i></p> <p><i>“I like the logo. It is a strong image. Secure and safe. It’s good”.</i></p> <p><i>“I like the shield because it’s like you have a knight protecting you”.</i></p> <p><i>“Actually, that logo is very apt. I like it.”</i></p>	<ol style="list-style-type: none"> 1. Safe and secure image 2. Strong message
<p><i>“I like the name; BullStop. When I heard the name, it resonates because cyberbullying and stopping cyberbullying. I just like the name. In short, I think its children friendly”.</i></p> <p><i>“The name is very clever. I got it immediately”.</i></p> <p><i>“It’s quite a smart name”.</i></p>	<ol style="list-style-type: none"> 1. Apt name 2. Clever

Theme	Good overall performance
Sample Quotes	Coded As
<p><i>“It correctly picked out the offensive sentences I typed”.</i></p> <p><i>“Even when I tried to be sneaky with a bad word. It got it. I wasn’t expecting that”.</i></p> <p><i>“I like that you can tick the checkboxes to improve the app.”</i></p>	<ol style="list-style-type: none"> 1. Accurate predictions 2. Consistent 3. Able to update predictions
<p><i>“It’s quite fast. I timed the message checker a few times, and it was like 1 or 2 seconds. That’s impressive”.</i></p> <p><i>“I don’t know why I keep expecting it to hang. I know that’s bad, but it didn’t at all”.</i></p> <p><i>“Going from screen to screen, it was quite smooth”.</i></p>	<ol style="list-style-type: none"> 1. Responsive 2. Exceeds expectations – ‘not hanging’ 3. Exceeds expectations – fast

Theme	User empowerment.
Sample Quotes	Coded As
<i>“I think it’s really good because if you want to see a text message and you don’t know what you’re in for, it might</i>	<ol style="list-style-type: none"> 1. Being in control 2. Feeling protected

APPENDIX E.12: Coding Tables for ‘Lab’-Based Evaluation Study Emergent Themes

<p><i>be something weird. It might just be something silly but still cheeky, and before you read that message, you can quickly scan the checkboxes, and you know what you're in for, and you know what you're supposed to do before you actually read the message".</i></p> <p><i>"It's good that with the app, I can control what kind of messages I receive".</i></p> <p><i>"I like that you can make contacts trusted or blocked. And that it can automatically block people if they are being offensive".</i></p> <p><i>"It's quite handy to manage friends from different social networks in one place".</i></p> <p><i>"The fact that you can adjust the app settings is very good. That you can control what type of messages are deleted".</i></p> <p><i>"I think having something like this gives the child some measure of control back, and I think that's key".</i></p> <p><i>"A thing like this, it's putting the young person in control, and giving them some trust, and saying, 'Actually, you know what? You've got this, you can take care of yourself'".</i></p> <p><i>"What we say to parents is instead of having this effect on your child, why don't you switch off the internet at home, take off the device from the child. But in a way that can sometimes contribute to the feeling of powerless in young people because now they have lost their phones and Internet because of this so it is key to demonstrate to them that they do have the power to fight this and I think this app can provide that".</i></p>	<p>3. Configurable</p>
<p><i>"I like the fact that there's no parent companion app".</i></p> <p><i>"I don't think any child will install an app if they know it allows their parents access to what they are doing on social media. So, I think the decision not to include parental monitoring in the app is the right one".</i></p> <p><i>"I think it's good to give children some control instead of sending parents copies of their messages. No child likes that, and they probably won't use the app".</i></p> <p><i>"I think not having a parents' or companion app, is very good and should be used a selling point of this app, and that's useful because you want young people to use it, otherwise, there's no point for it in the first place. Teenagers will never use it if they know there's parent supervision in the app".</i></p>	<ol style="list-style-type: none"> 1. Focussed on young people 2. Promoting independence

APPENDIX E.12: Coding Tables for 'Lab'-Based Evaluation Study Emergent Themes

Theme	Reflective and educative
Sample Quotes	Coded As
<p><i>"I think the one when I was sending a text, and then it checked for me because as much as I don't want to be hurt, I don't want to hurt people either. Also, if people don't want to get in trouble for sending something, they can use the app to tell them if what they are sending is bad".</i></p> <p><i>"I like the message checker a lot too. I think it is a key feature".</i></p> <p><i>"The charity list is a nice touch. That could become quite helpful".</i></p> <p><i>"The best thing for me about the app is I think it tells you to stop and think before sending a message, which I think young people struggle with. Some people might not have the intention to bully, but at the end of the day, what you have said has gone a long way to hurt or to bully another child. The fact that it has a stop and think section, which for any child who does not directly want to bully or does not have the intention of bullying, that is positive".</i></p>	<ol style="list-style-type: none"> 1. Reflective 2. Educative 3. Supportive 4. Access to additional help

Theme	Useful and unique.
Sample Quotes	Coded As
<p><i>"I'm not aware in my professional capacity of any software or app that particularly addresses cyberbullying. This is something we can recommend to our patients because we know that it can improve lifestyles".</i></p> <p><i>"I haven't come across anything similar to this".</i></p> <p><i>"It's pretty good, the BullStop app. It protects young people from the dangers of social media. I will tell my friends about it".</i></p>	<ol style="list-style-type: none"> 1. Lack of alternatives 2. Unique 3. Useful
<p><i>"Overall, I think that it's a very good app. I would definitely use it".</i></p> <p><i>"I'm not being cyberbullied, but if I am, I will definitely use the app".</i></p> <p><i>"Yes, I guess if I have been harassed. Maybe not so much now, but when I was younger, I can see myself installing something like this if it was available then".</i></p>	<ol style="list-style-type: none"> 1. Intention to use
<p><i>"For my kids, I would suggest to them that why don't you just filter all your messages through this app because you might be chatting with a friend and jokingly say something offensive or vice versa but an app like this just act as a filter, a shield".</i></p>	<ol style="list-style-type: none"> 1. Intention to recommend 2. Helpful 3. Fits with common advice 4. Safeguarding

APPENDIX E.12: Coding Tables for ‘Lab’-Based Evaluation Study Emergent Themes

<p><i>“I really like it, and I know a few people that I think this will be good for.”</i></p> <p><i>“If someone is being bullied online, I would definitely recommend this app to them, and if this tool is publicly available, it’s something the Police can suggest to people as a safeguarding tool”.</i></p> <p><i>“Definitely, I can recommend this [app] to them because it is putting the advice we provide into effect; block offensive contacts, review connection requests, etc., these are all the practical things we tell the parents to do”.</i></p> <p><i>“It is within our professional capacity to recommend apps like this because we know that these are tools that can help in their everyday lives”.</i></p>	
---	--

Theme	Suggested improvements
Sample Quotes	Coded As
<p><i>“I really like the app [...]. I think it’s great, but I don’t actually use Twitter. I am on Instagram and use WhatsApp a lot, and my friends are the same. I think if you can have Instagram and WhatsApp, I can see a lot of teenagers using this.”</i></p> <p><i>“If other social networks like Instagram can work with it, I think more people will want to use it”.</i></p> <p><i>“Working with different social networks will be key to this going viral. Sometimes bullies hop from one platform to another”.</i></p>	<p>1. Provide support for other online social networks</p>
<p><i>“I wasn’t sure about some icons”.</i></p>	<p>1. Confused by icons</p>
<p><i>“I would like if the message checker can be used in other apps like WhatsApp”.</i></p> <p><i>“It will be good if it can check as people are texting”.</i></p>	<p>1. Provide support for instant messaging and texting</p>
<p><i>“There are some messages that I sent that I expected like 4-5 checkboxes, but it picked 2.”</i></p> <p><i>“It’s quite good at picking abuse and bullying, but it didn’t get the threatening messages”.</i></p>	<p>1. Improve predictions precision</p>

APPENDIX E.13: Field-Based Evaluation Research Approval

From: Alshukry, Ali
Sent: 30 October 2019 08:56
To: Lumsden, Joanna (Jo) <J.LUMSDEN@aston.ac.uk>; Richards, Matthew <m.richards3@aston.ac.uk>
Cc: Salawu, Semiu D (Research Student) <salawusd@aston.ac.uk>
Subject: RE: Your advice....before you go!

Hi Jo,

I can confirm that the user testing is not research activity and therefore does not require research ethics review. This is more comparable to market research where the word “research” comes up but it is not meant for academic research. For this piece of work the academic research aspect is being done under research ethics. The next stage is to evaluate the product (in this case the app), therefore, as long as the legal aspects are covered, I do not see why you can’t go ahead with it.

You might want to think about liability insurance should the app damages someone’s device. Even if you covered this in the EULA you will still need some form of insurance should someone decides to take legal action for any other reason.

I have minor comments:

[Privacy policy:](#)

First paragraph – fifth line “collect” should be “collected”

Under “**Changes to This Privacy Policy**” there must an option to notify users of the change to go and change, otherwise this may be considered a GDPR breach as there is not active consent to the change.

Best wishes

Ali

Ali Alshukry

Research Integrity Manager



Research Integrity Office
Research and Knowledge Exchange
Aston University, Birmingham, B4 7ET, UK
0121 204 3738
Research_Governance@aston.ac.uk
www.aston.ac.uk

APPENDIX E.14: Field-Based Evaluation Study Questionnaire

1. How old are you?
 - 13 – 15
 - 16 – 18
 - Over 18

2. What is your gender?
 - Male
 - Female
 - Other non-binary
 - Prefer not to say

3. Which of the following do you regularly use (select all that apply)?
 - Twitter
 - Facebook
 - Myspace
 - Snapchat
 - WhatsApp
 - Instagram
 - Other (please specify): _____

4. How often do you use these social networks?
 - Several times a day
 - A few times a day
 - A few times a week
 - Rarely

5. Have you ever been cyberbullied?
 - Yes
 - No (Skip to question 11)

6. Who cyberbullied you?
 - A friend
 - Someone I know
 - Someone I don't know

7. When was this?
 - Within the last 3 months
 - Within the last 4-6 months
 - Within the last 7-12 months
 - Over a year ago

8. How did this cyberbullying occur? (select all that apply)
 - They sent offensive text messages to me
 - They sent offensive private messages to me on social media
 - They shared something offensive about me online publicly
 - Other (please explain): _____

APPENDIX E.14: Field-Based Evaluation Study Questionnaire

9. What did you do (select all that apply)?

- I told my parents
- I told a friend
- I just ignored them
- I told them to stop
- I didn't know what to do
- I didn't do anything because I was too scared/upset
- I retaliated (please explain): _____

10. Has this cyberbullying stopped?

- Yes
- No

11. Have you ever cyberbullied someone?

- Yes
- No (Skip to question 14)

12. Who did you cyberbully?

- A friend
- Someone I know
- Someone I don't know

13. How did this cyberbullying occur? (Select all that apply)

- I sent offensive text messages to them
- I posted something offensive about them online publicly
- I sent them offensive private messages on social media
- Other (please explain): _____

14. Please select all of following statements that you agree with:

- I don't know what cyberbullying is.
- Cyberbullying is not a big deal.
- Friends of mine have been cyberbullied.
- We've had cyberbullying incidents in my school.
- I have said things to others online that they didn't like but I don't consider it cyberbullying.
- Others have said mean things to or about me online, but I don't consider it cyberbullying.

15. How did you find out about the app?

- I saw a report about it (TV, newspaper or online)
- Through a charity organisation
- Word of mouth
- An online search (e.g. Google)
- A Play Store search
- Other (please specify) _____

APPENDIX E.14: Field-Based Evaluation Study Questionnaire

16. Do you think the app is a good idea?

- Yes
- No

17. Please tell us why you selected your answer in Question 16.

18. How easy was it for you to use the app?

- | Very Difficult | Difficult | Neither Easy Difficult | Easy | Very Easy |
|----------------|-----------|------------------------|------|-----------|
| 1 | 2 | 3 | 4 | 5 |

19. Did you take the app tour?

- Yes
- No (Skip to question 22)

20. Did you find the app tour useful?

- Yes
- No

21. Do you think you would have been able to confidently use the app without the tour?

- Yes
- No

22. In the period that you have been using the app, did it delete any offensive message?

- Yes
- No (Skip to question 26)

23. Approximately how many messages were deleted? _____

24. Did you have to correct the ticked checkboxes for the deleted messages (i.e., did you disagree with the app in terms of some of the messages it thought were cyberbullying)?

- Yes
- No

25. How many deleted messages did you do this for? _____

APPENDIX E.14: Field-Based Evaluation Study Questionnaire

26. How would you rate the app's ability to correctly delete offensive messages?

Very Bad	Bad	Average	Good	Very Good
1	2	3	4	5

27. Did you have to correct the ticked checkboxes for your received messages (i.e., did you disagree with the app in terms of some of the messages it allowed through but you thought were cyberbullying)?

- Yes
- No (Skip to question 29)

28. How many received messages did you correct? _____

29. The app flags sent messages in terms of how offensive it thinks they are: do you think this is a good idea?

- Yes
- No

30. Did you have to correct the ticked checkboxes for your sent messages (i.e., did you disagree with the app in terms of what it thought of some of your own messages)?

- Yes
- No (Skip to question 32)

31. How many sent messages did you correct? _____

32. How would you rate the app's ability in terms of correctly identifying **offensive** messages?

Very Bad	Bad	Average	Good	Very Good
1	2	3	4	5

33. How would you rate the app's ability in terms of correctly identifying **non-offensive** messages?

Very Bad	Bad	Average	Good	Very Good
1	2	3	4	5

34. Did you find the app icons self-explanatory?

- Yes
- No

APPENDIX E.14: Field-Based Evaluation Study Questionnaire

35. Do you feel that the app was well designed for people your age?

- Yes
 No

36. How would you rate the app in terms of how quickly it responds?

Very Bad	Bad	Average	Good	Very Good
1	2	3	4	5

37. Do you think the app will be useful for people your age?

- Yes
 No

38. What was your favourite feature of the app.?

39. What was your least favourite feature of the app?

40. How would you improve the app?

41. Would you continue using the app?

- Yes
 No

42. Please tell us why you selected your answer in Question 40.

APPENDIX E.14: Field-Based Evaluation Study Questionnaire

43. Overall, how would you rate the app?

Very Bad	Bad	Average	Good	Very Good
1	2	3	4	5

44. We would love to hear anything else you want to tell us about the app and what you think about it so that we can improve the app. Please leave us general comments below.