Approaches to Automated Detection of Cyberbullying: A Survey

Semiu Salawu, Yulan He, and Joanna Lumsden

Abstract— Research into cyberbullying detection has increased in recent years, due in part to the proliferation of cyberbullying across social media and its detrimental effect on young people. A growing body of work is emerging on automated approaches to cyberbullying detection. These approaches utilise machine learning and natural language processing techniques to identify the characteristics of a cyberbullying exchange and automatically detect cyberbullying by matching textual data to the identified traits. In this paper, we present a systematic review of published research (as identified via Scopus, ACM and IEEE Xplore bibliographic databases) on cyberbullying detection approaches. On the basis of our extensive literature review, we categorise existing approaches into 4 main classes, namely supervised learning, lexicon-based, rule-based, and mixed-initiative approaches. Supervised learning-based approaches typically use classifiers such as SVM and *Naïve Bayes* to develop predictive models for cyberbullying detection. Lexicon-based systems utilise word lists and use the presence of words within the lists to detect cyberbullying. Rule-based approaches match text to predefined rules to identify bullying, and mixed-initiatives approaches combine human-based reasoning with one or more of the aforementioned approaches. We found lack of labelled datasets and non-holistic consideration of cyberbullying by researchers when developing detection systems are two key challenges facing cyberbullying detection research. This paper essentially maps out the state-of-the-art in cyberbullying detection research and serves as a resource for researchers to determine where to best direct their future research efforts in this field.

Index Terms—Abuse and crime involving computers, data mining, machine learning, natural language processing, sentiment analysis, social networking

1 Introduction

Bullying is defined as intentional aggression carried out repeatedly by one individual or a group of individuals towards a person who is unable to easily defend him or herself (Olweus, 1993). Cyberbullying is, by extension, defined by Smith *et al.* (2008, pg. 376) as "an aggressive, intentional act carried out by a group or individual using electronic forms of contact, repeatedly or over time against a victim that cannot easily defend him or herself". Hinduja and Patchin (2009, pg. 5) define cyberbullying as "wilful and repeated harm inflicted through the use of computers, cell phones, and other electronic devices". Cyberbullying has been found to be quite prevalent on social media with as many as 54% of young people reportedly cyberbullied on Facebook (Ditch The Label, 2013). Zhang *et al.* (2016) found that neutralising processes (Sykes and Matza, 1957) play a significant role in why many young people engage in cyberbullying. They surmised that cyberbullies engage in such delinquent acts by rationalising their behaviours as valid and that the severity of possible sanctions does not deter.

There is substantial variation in the reported frequency for cyberbullying victimisation, with rates as low as 4%–5% reported by Olweus (2012) for the U.S.A. and rates as high as 35%–57% reported for mainland China (Zhou *et al.*, 2013). Patchin and Hinduja (2012) reported a frequency of about 20% amongst their survey of 4,400 students and found an average rate of 24% across existing studies. The EU Kids Online report (Livingston *et al.*, 2014) surmised that cyberbullying has now surpassed face-to-face bullying in the UK, with 12% of teenagers aged 9–16 years experiencing some form of cyberbullying victimisation as opposed to 9% for face-to-face bullying. This variation in the reported frequency of cyberbullying has been attributed to how cyberbullying has been defined by each study (Patchin and Hinduja, 2012) and the length of the intervening period between a cyberbullying incident and when victims were interviewed (Sabella *et al.*, 2013), with (perhaps unsurprisingly) the more recent victims of cyberbullying scoring higher on impacts and effects.

The detection of cyberbullying and online harassment is often formulated as a classification problem. Techniques typically used for document classification, topic detection, and sentiment analysis can be used to detect electronic bullying using characteristics of messages, senders, and the recipients. It should, however, be noted that cyberbullying detection is intrinsically more difficult than just detecting abusive content. Additional context may be required to prove that an individual abusive message is part of a sequence of online harassment directed at a user(s) for such a message to be labelled as cyberbullying. Thus, a tweet such as "*@username So you got drunk at a party and two people*

- *S. Salawu is with the Computer Science Research Group at Aston University, Birmingham, B4 7ET, UK. E-mail: salawusd@aston.ac.uk*
- *Y. He is with the System Analytics Research Institute at Aston University, Birmingham, B4 7ET, UK. E-mail: y.he9@aston.ac.uk*
- *J. Lumsden is with the Computer Science Research Group at Aston University, Birmingham, B4 7ET, UK. E-mail: j.lumsden@aston.ac.uk*

take advantage of you, that's not rape you're just a loose drunk slut #BiasedResults #Steubenville" can be easily classified as online harassment due its use of profanity ("slut") but requires additional context such as conversation history to determine if this is indeed bullying. Cyberbullying detection is inherently difficult due to the subjective nature of bullying. It extends beyond detecting negative sentiments or abusive content in a message as these tasks, on their own, do not necessarily mean that the message is in fact bullying. For example, a message such as "*I'm disgusted by what you said today and I never want to see you again*" is difficult to classify as bullying without understanding the larger context of the exchange, even though the message is clearly expressing very negative sentiments. Conversely, positively-expressed sentiments may disguise bullying if the intent is to express sarcasm.

We define cyberbullying detection as the identification of bullying actions (e.g., direct abuse, social exclusion, impersonation, sharing offensive materials) within an electronic communication medium and it comprises the following key tasks:

1. identification of individual bullying messages within a communication exchange;
2. and/or computing the severity of the bullying incident;
3. and/or identification of the roles inhabited by the individuals involved;
4. and/or 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).

We use this definition as part of our survey's inclusion criteria and only include studies that attempt one or more of the above tasks. In defining the roles identification task, we used the 8 roles identified by Xu *et al.* (2012a) as the superset of roles. These are of *bully*, *victim*, *bystander*, *assistant*, *defender*, *reporter, accuser,* and *reinforcer. Bystanders* are witnesses that do not intervene in a bullying incident. *Assistants* are co-perpetrators but not initiators. *Reinforcers*, while not directly involved in the bullying, encourage bullies and provide an impetus for continuation (e.g., laughing at the expense of victims). An *accuser* differs from a *reporter* by actively identifying *victim*s and *bullies*. Finally, *defenders* aid victims by coming to their aid. These roles encompass the various roles actors can inhabit during a cyberbullying incident and, as such, our sample includes studies that detected one or more of these roles. In fact, we did not find any study that attempted detecting roles outside of these 8 roles.

Given that the papers in our sample predominantly detect cyberbullying via textual features, our survey is therefore focussed on textual cyberbullying. Emerging areas such as detecting cyberbullying via image-, video-, and spoken-word analyses are not included, as our search did not discover papers attempting these tasks.

Nadali *et al.* (2013) and Kovacevic and Nikolic (2014) presented summaries of some existing research on cyberbullying detection and, while these 2 papers highlight some key research efforts in the area, they are neither exhaustive of extant literature nor do they provide detailed comparisons of the detection methods. This paper presents an in-depth review of the current state-of-the-art in cyberbullying detection and provides a unique contribution to cyberbullying research by identifying, categorising, and reviewing current and existing work in the field. To our knowledge, this survey effort is unique within the field of cyberbullying detection and forms the first phase of our research into the creation of a mobile tool for the prevention of cyberbullying on social media.

The following section details our survey methods and the results of our review and in Section 3 we discuss observations from our sample.

2 Survey Methods and Results

In this section, we present an overview of our search strategy and the results of our methodical survey of the literature. We discuss the search conducted to locate the studies reviewed and the data abstractions used to categorise the studies. Finally, we present 4 tables illustrating all the studies along with the key characteristics used to discuss the studies in the rest of the paper.

2.1 Data Search and Selection

An electronic literature search was conducted across Scopus, the ACM Digital Library, and the IEEE Xplore digital library. The main search strategy was the discovery of academic literature relevant to the theme "automated detection of electronic bullying, anti-social behaviour and harassment" using the following query phrases without any publication year filter applied:

"cyber-bull* or cyberbull* detection", "detecting cyberbull* or cyberbull*", "electronic or online bullying detection", "detecting electronic or online bullying, cyber-bull*" or "cyberbull* prevention tool", "cyber-bull* or cyberbull* prevention software", "cyber-bull* or cyberbull* software", "anti cyber-bull* or anti cyberbull*" or "anti-cyber-bull* or anti-cyber-bull*" or "anticyberbull* or anticyberbull*", "detecting electronic or online harassment".

A citation trail was performed on the discovered papers using the papers' references as a starting point and a total of 89 academic papers was discovered as a result of the search. The papers were initially assessed for relevance via a review of their titles, abstract, and concluding arguments: 18 papers were not considered relevant to the survey and so were removed. The full text of the remaining papers was reviewed and papers whose primary focus did not include any of the 4 cyberbullying detection tasks we identified in Section 1 were discounted. This led to the removal of a further 18 papers. These included papers that dealt with themes such as youth violence involvement detection (Sigel and Harpin, 2013), story matching to identify distressed teens (Dinakar *et al.*, 2012b; Macbeth *et al.*, 2013), and cyberbully prevention policies (Al Mazari, 2013). To eliminate the effects of language on cyberbully detection when comparing the reviewed studies, we excluded papers using non-English corpora; thus a further 7 papers

were excluded. These included papers such as Ptazynski *et al.* (2010a; b), Honjo *et al.* (2011), Nitta *et al.* (2013), Li and Tagami (2014), Margono *et al.* (2014) and Van Hee *et al.* (2015) which were removed as they used non-English corpora.

The remaining 46 papers were included in the final list of papers examined by this study.

2.2 Data Abstraction

For the included papers, we performed data abstraction using characteristics such as detection tasks performed, data sources, the size and availability of the datasets, detection techniques, annotation judgement, features extracted, external resources used, and pre-processing steps. We used the total number of documents (i.e., messages, posts, comments, etc.) as a measure of the data size as opposed to using other metrics such as the number of users or threads in a dataset; thus, a sample containing 50 messages generated by 70 users was assigned 50 as the data size value.

2.3 Dimensions of Characterization

Tables 1 and 2 present a summary of key information abstracted from the reviewed studies. Table 1 provides a quick overview of approach categories and detection tasks for each of the 46 papers. Table 2 presents additional information about the studies, such as features and techniques used, pre-processing steps performed, and any external resources used (e.g., WordNet[1], urbandictionary[2], etc.). Table 3 presents details (where available) of the datasets used by the papers. Finally, the best available results per detection tasks for each corpus category are presented in Table 4.

TABLE 1: STUDIES, TASKS PERFORMED AND APPROACH CATEGORIES

Study	Tasks	Approach Category
Mahmud *et al.*, 2008	Binary Classification	Rule-based
Yin *et al.*, 2009	Binary Classification	Supervised Learning
Bosse and Stam, 2011	Role Identification	Other (BDI Agents)
Dinakar *et al.*, 2011	Binary Classification	Supervised Learning
Sanchez and Kumar, 2011	Binary Classification, Role Identification	Supervised Learning
Serra and Venter, 2011	Binary Classification	Rule-based
Burn-Thorton and Burman, 2012	Binary Classification	Supervised Learning
Chen *et al.*, 2012	Cyberbullying Severity, Role Identification	Rule-based
Dadvar and De Jong, 2012	Binary Classification	Supervised Learning
Dadvar *et al.*, 2012a	Binary Classification	Supervised Learning
Dadvar *et al.*, 2012b	Binary Classification, Classification of follow-on events	Supervised Learning
Dinakar *et al.*, 2012a	Binary Classification	Mixed Initiative
Mancilla-Caceres *et al.*, 2012	Role Identification	Other (Human Judgement)
Nahar *et al.*, 2012	Binary Classification, Role Identification	Supervised Learning
Perez *et al.*, 2012	Cyberbullying Severity	Lexicon-Based
Sood and Churchill, 2012a	Binary Classification	Supervised Learning
Sood and Churchill, 2012b	Binary Classification	Supervised Learning
Xu *et al.*, 2012a	Binary Classification, Role Identification	Supervised Learning
Xu *et al.*, 2012b	Sentiment Analysis	Supervised Learning
Dadvar *et al.*, 2013a	Cyberbullying Severity	Mixed Initiative
Dadvar *et al.*, 2013b	Cyberbullying Severity	Supervised Learning
Kontostathis, 2013	Binary classification	Supervised Learning
Munezero, 2013	Binary Classification	Supervised Learning
Nahar *et al.*, 2013	Binary Classification, Role Identification	Supervised Learning
Sheeba and Vivekanandan, 2013	Binary classification	Supervised Learning
Bretschneider *et al.*, 2014	Binary Classification	Rule-based, Lexicon-Based
Dadvar *et al.*, 2014	Role Identification	Mixed Initiative
Del Bosque and Garza, 2014	Cyberbullying Severity	Supervised Learning
Fahrnberger *et al*, 2014	Binary Classification	Lexicon-Based
Galán-García *et al.*, 2014	Role Identification	Supervised Learning
Huang et al., 2014	Binary Classification	Supervised Learning
Nahar *et al.*, 2014	Binary Classification	Semi-Supervised Learning
Munezero, 2014	Binary Classification	Supervised Learning
Parime and Suri, 2014	Binary Classification	Supervised Learning
Potha and Maragoudakis, 2014	Binary Classification, Cyberbullying Severity	Supervised Learning
Chavan and Shylaja, 2015	Binary Classification	Supervised Learning
Hosseinmardi et al, 2015	Binary Classification	Supervised Learning
Mancilla-Caceres *et al.*, 2015	Role Identification	Other (Human Judgement)
Mangaonkar et al., 2015	Binary Classification	Supervised Learning
NaliniPriya and Asswini, 2015	Binary Classification	Supervised Learning
Nandhini and Sheeba, 2015a	Binary Classification	Supervised Learning
Nandhini and Sheeba, 2015b	Binary Classification	Supervised Learning

[1] wordnet.princeton.edu

[2] www.urbandictionary.com

Rafiq et al, 2015	Binary Classification	Supervised Learning
Squicciarini et al, 2015	Role Identification, Classification of follow-on events	Supervised Learning
Zhao and Mao, 2016	Binary Classification	Supervised Learning
Zhao et al., 2016	Binary Classification	Supervised Learning

Our survey revealed binary classification as the most common task performed in cyberbullying detection. In this regard, bullying messages are considered members of a "bullying" class and all other documents belong to the "other" or "non-bullying" class. The key task then is the identification of documents that possess the core attributes of the "bullying" class. Out of the 46 studies reviewed, 34 performed binary classification either as the sole detection task or in combination with other tasks. This classification of messages is often facilitated by sentiment analysis using emotive wordlists, supervised learning, and lexicon-based systems. Studies such as Yin *et al.* (2009), Dinakar *et al.* (2011), Xu *et al.* (2012a), and Rafiq *et al.* (2015) performed sentiment analysis using supervised-learning techniques. Others such as Burn-Thorton and Burman (2012), Kontostathis *et al.* (2013), Nahar *et al.* (2013; 2014), Munezero *et al.* (2014), Nandhini and Sheeba (2015a;b), and Zhao *et al.* (2016), while also implementing binary classification, did not perform the message classification via sentiment analysis. Interestingly, Xu *et al.* (2012a) is the only instance we found whereby sentiment analysis is performed not for the purpose of binary classification but to understand the emotions expressed in what they term "*bully traces*", which are tweets containing any of the words "*bully*", "*bullied*" and "*bullying*" (i.e., tweets containing bullying references or reportage – e.g., "*I saw a girl got bullied at school today #bullyingisnotcool*"). Role identification is the next most performed task (11 papers), featuring heavily in studies such as Sanchez and Kumar (2011), Chen *et al.* (2012), Dadvar *et al.* (2014), and Galán-García et al. (2014).

Determining the severity of cyberbullying by computing a score indicative of the bullying severity of messages and/or sender is performed by studies such as Chen *et al.* (2012), Perez *et al.* (2012), Dadvar *et al.* (2013a), Del Bosque and Garza (2014), and Potha and Maragoudakis (2014). Dadvar *et al.* (2012b) and Squicciarini *et al.* (2015) were the only studies we found that proposed the relatively novel task of detecting and classifying the events that occur *after* a cyberbullying incident.

While cyberbullying occurs across various forms of electronic media – such as SMS (Short Messaging Service), MMS (Multimedia Messaging Service), email, forums, chat rooms – and social media platforms like Facebook, Twitter, YouTube and SnapChat, social media was the main source of data for many of the studies reviewed. This can be attributed to the availability of social media data which is often freely accessible in the public domain; emails, SMS, MMS and chat rooms are, in contrast, very personal means of communication and, as such, communications via these media are less likely to be publicly available.

Twitter and MySpace are the most common data sources. Twitter is used in many studies including Sanchez and Kumar (2011), Xu *et al.* (2012a; b), Huang *et al.* (2014), Galán-García et al. (2014), and Zhao *et al.* (2016). MySpace is used by Yin *et al.* (2009), Parime and Suri (2014), Nandhini and Sheeba (2015a; b), and Squicciarini *et al.* (2015) amongst others. YouTube is in second place with Dinakar *et al.* (2011), Chen *et al.* (2012), Dadvar *et al.* (2013a; b; 2014) using corpora that included YouTube data. Burn-Thorton and Burman (2012) is the only paper in our sample that uses an email corpus. 14 papers publicly shared their datasets: 9 of these make use of the Barcelona Media dataset (a publicly available dataset of social media data) and the remaining 5 papers sourced the corpus themselves.

With supervised-learning methods proving popular amongst the reviewed studies (34 papers), the means by which judgements on annotated data were arrived at is of interest. Traditional means of labelling data using annotators or by the researchers themselves still proved to be popular, with 25 studies employing annotators, experts, or researchers to label data. Crowd-sourcing annotators is also gaining traction within the cyberbullying research community, with studies such as Sanchez and Kumar (2011), Kontostathis *et al.* (2013) and Hosseinmardi *et al.* (2015) using crowdsourcing services like Amazon Mechanical Turk (MTurk) and CrowdFlower to label data. Given the ease, relative low cost, and huge time savings of crowdsourcing, we expected to find higher utilisation of crowdsourcing services amongst the studies but perhaps researchers' need to ensure high-quality annotated data currently presents a barrier that crowdsourcing services will need to overcome in order to become more widely used. Interestingly only 3 papers (Dinakar *et al.,* 2011; 2012a; Rafiq *et al., 2015*) employed experts to annotate data. This is surprising since a natural assumption would be that the use of experts for annotation likely presents the best chance of achieving quality, labelled data. A possible reason for this low utilisation of experts for labelling data could be the subjective nature of bullying, a consequence of which could be that researchers' and experts' views on cyberbullying may differ greatly. Thus, researchers adopting a specific definition of cyberbullying would naturally want the annotators to be guided by this definition.

2.3.1 Features Used for Cyberbullying Detection

We broadly categorise features used across the studies into 4 main groups, namely content-, sentiment-, user- and network-based features. We define content-based features as the extractable lexical items of a document such as keywords, profanity, pronouns, and punctuations. Emotion-based features are those features that are indicative of emotive content; they are generally keywords, phrases and symbols (e.g., emoticons) that can be used to determine the sentiments expressed in a document. User-based features are those characteristics of a user's profile that can be used to make a judgement on the role played by the user in an electronic exchange and include age, gender, and sexual orientation. Finally, network-based features are usage metrics that can be extracted from the online social network and include items such as number of friends, number of followers, frequency of posting, etc.

TABLE 2: STUDIES AND THE FEATURES, CLASSIFIERS, PRE-PROCESSING AND EXTERNAL RESOURCES USED.

Study	Features				Classifiers	Pre-processing	External Resource Used
	Content	Sentiment	User	Network			
Mahmud *et al.,* 2008	Cyberbully keywords					Extraction of subject/event, Tokenization	OpenNLP, Stanford Parser
Yin *et al.,* 2009	TFIDF, Pronouns, N-gram				SVM	Tokenization, Stemming	
Bosse and Stam, 2011			User behaviours		BDI Agents		
Dinakar *et al.,* 2011	Cyberbully keywords, TFIDF, Profanity, N-gram	Sentiments			Naïve Bayes, SVM, J48, JRip	Stop words removal, Stemming, Removal of unimportant character sequence	
Sanchez and Kumar, 2011	Cyberbully keywords	Sentiments			Naïve Bayes	Keywords extraction	
Serra and Venter, 2011			Age	Time online	Neural Networks		
Burn-Thorton and Burman, 2012	Text				kNN		
Chen *et al.,* 2012	Profanity		Writing style		Naïve Bayes, SVM	Spelling and grammar correction	Stanford Parser
Dadvar and De Jong, 2012	TFIDF, Pronouns, Profanity		Gender		SVM		
Dadvar *et al.*, 2012a	TFIDF, Profanity Pronouns		Gender		SVM		noswearing.com
Dadvar *et al.*, 2012b	TFIDF		Age, Gender		SVM		
Dinakar *et al.*, 2012a	Cyberbully keywords, TFIDF, Profanity, N-gram	Sentiments			Naïve Bayes, SVM, J48, JRip		
Mancilla-Caceres *et al.,* 2012				User Interactions			
Nahar *et al.,* 2012	N-gram	Sentiments		Network nodes	SVM		
Perez *et al.,* 2012	Cyberbully keywords, Profanity					Tokenization, Removal of unwanted characters, Characters substitution	
Sood and Churchill, 2012a	TFIDF, N-gram, Profanity, Levenshtein Distance				SVM	Stemming	phorum.com, noswearing.com
Sood and Churchill, 2012b	N-gram, Profanity, Levenshtein Distance				SVM	Stemming	phorum.com, noswearing.com
Xu *et al.*, 2012a	N-gram	Sentiments			Naïve Bayes, SVM, Logistic Regression, LDA, Conditional Random Fields (CRF)	Removal of unimportant character sequence	
Xu *et al.*, 2012b	N-gram	Sentiments			Naïve Bayes, SVM, Logistic Regression, LDA		

Dadvar *et al.*, 2013a	Comment length, Profanity, Spelling, Pronouns		Age	Membership duration, uploads, subscriptions, comments	MCES		
Dadvar *et al.*, 2013b	Cyberbully keywords, Pronouns, Profanity, Capitalisation, Emoticons, N-gram, Message length		Age	User's activity history	SVM	Stemming, Stop words removal	noswearing.com
Kontostathis, 2013	Profanity				EDLSI	Case conversion, characters removal	
Munezero, 2013	N-gram				Naïve Bayes, SVM, J48	Tokenization, Stemming, Stop words removal	
Nahar *et al.,* 2013	TFIDF, Pronouns, Profanity				SVM, LDA, HITS		
Sheeba and Vivekanandan, 2013	TFIDF	Sentiments			Maximum Entropy, Fuzzy Systems	Stemming, Stop words removal	WordNet, SentiWordNet
Bretschneider *et al.,* 2014	Profanity, Pronouns					Tokenization, POS tagging, Spelling and grammar correction	http://www.infochimps.com/collections/moby-project-word-lists
Dadvar *et al.,* 2014	TFIDF, Comment length, Profanity, spelling, Pronouns		Age	Membership duration, uploads, subscriptions, comments	Naïve Bayes, SVM, C.45, MCES		
Del Bosque and Garza, 2014	Document length, Profanity, Pronouns	Sentiments			Multi-Layer Perceptron (MLP) Neural Network		noswearing.com, ANEW, SentiWordNet
Fahrnberger *et al,* 2014	Profanity	Sentiments					
Huang et al., 2014	Profanity, Capitalisation, Punctuation, Emoticons, POS Tags			Ego Networks	J48, Naïve Bayes, SMO, ZeroR		
Nahar *et al.,* 2014	Pronouns, Profanity, Capitalisation, Special Characters	Sentiments	Age	Gender	Radom Forrest, K-FSVM, Naïve Bayes, Logistic Regression		
Munezero, 2014	N-gram	Sentiments			Naïve Bayes, SVM, J48	Tokenization, Stemming, Stop words removal	SentiStrength, WordNetAffect
Parime and Suri, 2014					SVM (Linear)	Tokenization, Stop words removal, Stemming	
Potha and Maragoudakis, 2014	N-grams (BoW), TFIDF, Term frequency, Term occurrence				SVM, Multi-Layer Perceptron (MLP) Neural Network	Tokenisation, Stop words removal, Case conversion	
Chavan and Shylaja, 2015	N-gram, Word count, TFIDF, Pronouns, Skip-grams				SVM, Logistic Regression	Removal of unwanted characters, Spelling correction	

Galán-García *et al.,* 2014	Tweet, Language			Time of Posting, Location, Twitter client	SVM-PolyKernel, J48, SVM-Normalized-PolyKernel , Random-Forest		
Hosseinmardi et al, 2015	N-grams, Image features			followers, following, likes	Naïve Bayes, SVM		
Mancilla-Caceres *et al.,* 2015				Players Interactions			
Mangaonkar et al., 2015	N-gram				Naive Bayes (NB), Logistic regression, SVM	Tokenization	
NaliniPriya and Asswini, 2015	Profanity			Ego Networks			
Nandhini and Sheeba, 2015a	Word count, Nouns, Pronouns, Adjectives				Naïve Bayes	Stop words removal, Removal of unwanted characters, POS tagging	
Nandhini and Sheeba, 2015b	Profanity, Cyberbully keywords, BoW				Naïve Bayes	Stop words removal, Removal of unwanted characters, POS tagging	
Rafiq et al, 2015	N-grams	Sentiments		Followers, following, media uploads, likes, comments, views	Naïve Bayes, AdaBoost, Decision-Tree, RandomForest		
Squicciarini et al, 2015	Profanity, Length of post, Pronouns, BoW	Sentiments	Age, Gender,	Elapsed time between comments, friends, Centrality in social network	C4.5 Decision Tree		
Zhao and Mao, 2016	Profanity, Cyberbully keywords, BoW				SVM (Linear)		
Zhao et al., 2016	EBoW				SVM (Linear)	Tokenisation	

2.3.1.1 Content-based Features

We group features such as cyberbullying keywords, profanity, pronouns, n-grams, Bags-of-words (BoW), Term Frequency Inverse Document Frequency (TFIDF), document length, and spelling as content-based features.

Content-based features are overwhelmingly used across our sample, with as many as 41 papers utilising content-based features. As cyberbullying messages are often abusive and insulting in nature, it is not surprising that profanity was found to be the most used content-based feature across the reviewed studies, with 22 papers using the presence of profanity in text as an indicator for cyberbullying. Studies such as Dinakar *et al.* (2011), Perez *et al.* (2012), Kontostathis *et al.* (2013), Nahar *et al.* (2013) and Bretschneider *et al.* (2014), created profanity lexicons using word lists compiled by the researchers or sourced from external libraries such as noswearing.com[3] and urbandictionary.com. By equating the presence of profanity to cyberbullying, the use of profanity lexicons alone fails to consider other key aspects of cyberbullying such as repetitiveness and the presence of a power differential. Rafiq *et al.* (2015) similarly cautioned against the use of profanity as the only feature for cyberbullying detection and argued that not all use of profanity and cyber-aggression constitutes bullying. Studies such as Nahar *et al.* (2013), Dadvar *et al.* (2014), Bretschneider *et al.* (2014) and Nahar *et al.* (2013) incorporated other features such as pronouns in close proximity to profanity, since such personalised abusive content is potentially more indicative of cyberbullying than abusive terms on their own. For example, the phase "*the f**king train was delayed again*" is definitely not cyberbullying although it contained profanity but "*you f**king idiot*" could be. While this is an improvement, the *pronoun + profanity* feature still suffers the same shortcomings as using profane terms alone.

Dinakar *et al.* (2011), often cited for the performance gain achieved by their label-specific binary classifiers over multi-class classifiers, achieved this improved performance by using domain-specific content features learned from training classifiers on a set of messages clustered on sensitive topics such as race, culture, sexuality, and intelligence to then detect bullying messages within each cluster.

While Yin *et al.* (2009) did not find n-grams very effective in their experiments, its use as a detection feature is still relatively popular amongst studies, including Dinakar *et al.* (2011), Xu *et al.* (2012a; b), Sood and Churchill (2012a; b), and Munezero *et al.* (2014). As TFIDF provides a measure of a word's importance to a document within a collection of documents, it can sometimes provide better results than using n-grams in isolation (Yin *et al.*, 2009). It is, therefore, often used alongside n-gram and other features to improve detection performance, as can be seen in the works of Yin *et al.* (2009), Dinakar *et al.* (2011), Dadvar and De Jong (2012), Sood and Churchill (2012a), and Nahar *et al.* (2013).

Of the 41 studies using content-based features, 5 checked for the presence of cyberbullying keywords as part of the detection process. By cyberbullying keywords, we refer to non-profane words the use of which can indicate the presence of cyberbullying. These often are words associated with themes such as race, physical appearance, gender, and sexuality. As far back as the earliest study we discovered (i.e., Mahmud *et al.*, 2008), cyberbullying keywords have been used as detection features and this trend has continued with later studies such as Dinakar *et al.* (2011), Sanchez and Kumar (2011), Perez *et al.* (2012) and Dadvar *et al.* (2013b). These studies created lexicons composed of words so selected because their presence within a message or a post connotes a high likelihood of cyberbullying. For example, both Dinakar *et al.* (2011) identified themes such as race, culture, sexuality, physical appearance, and intelligence as common bullying topics and used a lexicon of words associated with these themes as features, while Sanchez and Kumar (2011) concentrated on homophobic slurs such as "*gay*", "*queer*", "*homo*" and "*dyke*" as keywords.

Other content-based features we found used by studies include document length (Dadvar *et al.*, 2013a; 2014), word capitalisation (Dadvar *et al.*, 2013b; Nahar *et al.*, 2014), spelling (Dadvar *et al.*, 2013a; 2014), and the use of special characters (Nahar *et al.*, 2014).

2.3.1.2 Sentiment-Based Features

Sentiment or emotion analysis has been 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). Within the field of cyberbullying detection, sentiment analysis is often combined with features like TFIDF and pronoun usage to improve the performance of the detection system. This is due to the fact that, while strong emotions can often be an indicator of bullying, they are rarely sufficient on their own to accurately identify a bullying episode. For example, a sarcastic sentence such as *"I'm in love with your big nose"* that scores high on positive emotions may also constitute bullying and would require additional methods to identify the phrase "*big nose*" as an instance of a potentially negative remark about an individual's physical appearance. If, however, within the same sentence "*nose*" is replaced by "*eyes*", this may very well be a declaration of affection or genuine admiration.

We discovered 13 papers using emotion-based features within our survey. Typically, this involved the use of emotive keywords to perform sentiment analysis on the corpus and then using the discovered sentiment as an input to the detection process. With the exception of Nahar *et al.* (2012), who used Probabilistic Latent Semantic Analysis (PLSA) to extract sentiment features from labelled bullying posts, all the studies in this group used a lexicon of emotive words to detect the polarity (negative, positive, or neutral) of sentiments expressed within the documents. The emotive words are often based on sources such as WordNet and its variations.

Xu *et al.* (2012a) extracted "*bully traces*" via the Twitter Streaming API and identified the roles played by people referenced within the tweets. By reviewing the extracted

[3] www.noswearing.com

tweets, they detected seven different types of emotions in the tweets. These are anger, embarrassment, empathy, fear, pride, relief, and sadness.

In their subsequent work, Xu *et al.* (2012b) used labelled data from Wikipedia to train an *SVM* classifier to detect the emotions expressed in the tweets. The study found only 6% of the tweets contained any of the seven emotions; within this 6%, fear was expressed in half of the tweets, sadness in 19%, anger (18%) and relief (11%). Further analysis of the tweets revealed, however, that fear is often expressed jokingly (e.g., "*oooh I'm so scared*"), thus providing further evidence that a detection system based on sentiments alone cannot always accurately distinguish between genuine emotions and those sarcastically expressed. This is in agreement with Dinakar *et al's.* (2011) discovery that bullying involving deliberate abuse and profanity were much easier to detect that those containing sarcasm and euphemism.

Munezero *et al.* (2014) expanded the method proposed in Munezero *et al.* (2013) by introducing two emotion-based features directed at exploiting the emotional context of a post. The first emotion feature used an ontology of emotions and emotive words based on WordNetAffect (Strapparava and Valitutti, 2004) to determine the emotions expressed within text. The second feature used SentiStrength (Thelwall *et al.,* 2010) to calculate the emotional strength of a piece of text. The inclusion of these emotion-based features improved the detection process in the majority of the experiments conducted although, when compared to the results obtained in their earlier experiments using content-based features alone (Munezero *et al.*, 2013), these improvements were not significant. Interestingly, using the emotion-based features alone consistently yielded the lowest performance across several experiments.

2.3.1.3 User-Based Features

Alongside content-based and emotion-based features, researchers have explored incorporating user-related features into cyberbullying detection systems. These include features like age, gender, sexual orientation, and race. Our survey revealed age and gender to be the most commonly used user-based features, with papers including Serra and Venter (2011), Nahar *et al.* (2014) and Squicciarini *et al.* (2015) using either or both as features.

Dadvar and De Jong (2012) and Dadvar *et al.* (2012a) used the TFIDF of profane words and pronouns as features in a gender-specific corpus of MySpace posts to train an *SVM* classifier. They found cyberbullying detection was significantly improved by the inclusion of gender-specific features when compared against results obtained using the same classifier trained on a non-segregated dataset. While the improvements demonstrated by the study provide encouragement for the incorporation of gender features in online bullying detection, it should be noted that gender (and any other user-supplied) information on social media can be easily falsified. As such, any method that makes use of user-supplied information will greatly benefit from reliable means of validating such data – for example, a forensic linguistic module could be used to assign a "truth score" to age and gender information supplied by a user.

Serra and Venter (2011) examined cyberbullying via mobile phones and devised a pre-emptive approach for combating cyberbullying using a rule-based framework. The system assigns a risk profile to individuals based on the user's age and mobile phone usage pattern.

Once the user has been matched to a rule, the tool can then initiate an appropriate action, such as blocking access to the Internet or sending an alert to parents. It can be argued that this is quite a simplistic approach to detecting cyberbullying and that it is unlikely to detect cyberbullying episodes as it only assesses usage patterns at a rudimentary level. While heavy Internet usage has been identified as a cyberbullying risk factor in young children (Mishna *et al.*, 2012), a generalised rule flagging any high Internet usage does not take into account instances of legitimate need for prolonged Internet use (e.g., school work) or the type of activities in which the child is engaged (e.g., a child may use IM to stay in touch with parents). In addition, the proposed tool is incapable of determining if the messages exchanged constitute bullying or not; hence, a young user that only uses the Internet for a few hours a week but receives abusive messages within these hours will not be flagged as being at high risk. The system is also incapable of linking Internet usage by the same user across multiple devices (e.g., laptop or tablet), as is often the case nowadays.

Chen *et al.* (2012) also incorporated users' writing styles and conversation history as features in the development of their Lexical Syntactic Framework (LSF) and compute an offensiveness score for the user based on these features.

2.3.1.4 Network-Based Features

With the huge popularity of social media, including its status as the predominant source of data for cyberbullying detection research, it is not surprising that network data such as number of friends, uploads, likes and so on is increasingly being used as features in detection systems.

Serra and Venter (2011) is the earliest study in our sample using network-based features; they used total time present online using a mobile phone as a feature in their detection method. Nahar *et al.* (2012), Huang *et al.* (2014), and NaliniPriya and Asswini (2015) used ego networks as features to improve detection. NaliniPriya and Asswini (2015) used the ego network to compute temporal changes in the relationships between users, and uses the detected changes within the detection process. Huang *et al.* (2014) discovered that the risk of cyberbullying is decreased in ego networks with many people and high interconnectivity (probably because in such networks there is likely to be increased social support for potential victims) but that a higher number of messages exchanged between users indicate a higher likelihood for cyberbullying.

Dadvar *et al.* used membership duration, number of uploads, subscriptions, and comments posted as features in (2013a; 2014) and activity history in (2013b), alongside user-based and content-based features to achieve improved detection compared to experiments without network-based features. Mancilla-Caceres *et al.* (2012; 2015) used players' interactions within a social computer game

as the only feature to detect cyberbullying and Galán-García *et al.* (2014) were able to narrow down the likely perpetrators behind a twitter trolling profile by using network-based features like time of posting, location, and Twitter client.

In Squicciarini *et al.* (2015), the authors used the elapsed time between comments to measure the influence of cyberbullies on other users and the proliferation of bullying across a social network. Followers' numbers on social networks were used as features by both Rafiq *et al.* (2015) and Hosseinmardi *et al.* (2015) with Rafiq *et al.* (2015) supplementing this with other network-based features such as media uploads, likes, comments and views.

2.3.2 Pre-Processing of Data

It is not uncommon for the content provided as input to natural language processing tasks to first undergo a number of pre-processing phases. This is often performed to reduce noise within the data, thereby improving overall accuracy. Pre-processing, however, can be a double-edged sword as useful context can be lost during the process. For example, a common pre-processing step is the conversion of uppercase words to lowercase; this action may unintentionally result in the loss of context as capitalisation is often used to denote shouting in textual communication.

Pre-processing is performed by 22 studies in our sample, with tokenization and stemming used more than any other pre-processing steps. Stemming is often performed on a corpus when TFIDF, n-gram, and BoW are used as features. This is a logical endeavour since, by reducing words to their stems, the frequencies of such words are collapsed into a single value for the stem, thus accentuating the importance of such words within the corpus. Tokenization is often employed to break sentences and phrases into a sequence of characters (often individual words) and performed to enable a document to be represented as a function of its words.

The other key pre-processing tasks we discovered from the studies included in our survey are stopwords removal, character removal/substitution, grammar and spelling correction. Stopwords removal is aimed at eliminating common words that appear to be of little value to the domain in question. While it can reduce noise in the data, it can also inadvertently delete important terms. A better method could be to first determine if stopwords are used within named entities or commonly used phrases (using named-entity detection or phrase chunker) before removal.

Dinakar *et al.* (2011), Chen *et al.* (2012), Perez *et al.* (2012), Kontostathis *et al.* (2013), and Bretschneider *et al.* (2014) improved the quality of the data by removing repeated characters (e.g., *heeeeeeey*) and correcting spelling and grammar. Kontostathis *et al.* (2013) found that some of these steps can, however, also corrupt the data; for example, they discovered that legitimate words were also being affected – e.g., "*good*" became "*god*" – changing the meaning of the entire sentence. It can also be argued that excessive repetition of characters in words can often be intentional for emphasis (e.g., *you are such a biiiiiig idiot*) rather than a misspelling; thus, interpreting such as additional emphasis may be a better tactic than auto correction.

2.3.3 Cyberbullying Detection Techniques

The vast majority of the papers included in our survey used supervised learning techniques to detect cyberbullying, with Yin *et al.* (2009) being the earliest study we found using this technique. Based on the key approaches they employed, we classify the other techniques in use by the studies we reviewed as lexicon-based systems, rule-based systems, and mixed-initiative systems; we include an 'Other' category for approaches that do not fit into any of the above-listed classes.

2.3.3.1 Supervised Learning Approaches

In Yin *et al.* (2009), the authors analysed posts and comments from three different social websites (Slashdot[4], Kongregate[5] and MySpace[6]). Discovering that the percentage of harassment posts within a corpus is very small, they therefore hypothesised that a harassment post will appear significantly different from its neighbouring posts. On this basis, they introduced a document's immediate neighbourhood of k posts (k = 3) as a feature to an *SVM* (Support Vector Machine) classifier and saw an improved performance in the classification output compared to experiments without the neighbourhood feature.

Dinakar *et al.* (2011) performed two sets of experiments, first training 4 classifiers – *Naïve Bayes, JRip, J48,* and *SVM* – on a set of messages clustered by themes and then on an amalgamated set of all messages, and found performance much improved on individual clusters over the combined set. Thus, by first training on messages clustered on themes such as racism, culture, sexuality, and intelligence, the classifiers were able to learn better features to then identify bullying messages within each cluster. Essentially, cyberbullying detection was decomposed into a two-stage process with the first stage focused on clustering messages based on topics relevant to cyberbullying, followed by a second stage aimed at detecting profanity and negativity in the content. Dadvar and De Jong (2012) and Dadvar *et al.* (2012a) adopted a similar approach by training an *SVM* classifier on MySpace posts segregated by the writers' gender. They found cyberbullying detection was significantly improved on the gender-segregated posts when compared against results obtained when the same classifier was trained on a non-segregated dataset.

Following on from this work, Dadvar *et al.* (2012b) theorised that a content-based approach alone is not sufficient to detect bullying content, and advocated an approach that incorporates the impact felt by the receiver in order to accurately determine the severity of the bullying episode. This can be done by analysing a receiver's reply or follow-on actions within the same or another environment. For example, a victim may change his/her status on Facebook after receiving bullying text messages on a mobile phone and such status updates can be classified to determine the vic-

[4] slashdot.org
[5] kongregate.com
[6] myspace.com

tim's emotional state. Garlan-Garcia *et al.* (2013) also attempted to determine the victim's emotional state by using a sequential set of features to train an *SVM* classifier on a dataset of YouTube comments. They discovered that the most effective words for classification were profane words relating to race and sexuality. This is in agreement with Dinakar *et al.* (2011) who found it easier to detect cyberbullying after first segregating messages based on topics such as race, sexuality, and physical appearance

Nahar *et al.* (2014) also experimented with clustering messages as part of the detection process. They used *Kernel-based Fuzzy C-Means (K-FCM)* to cluster the data by evaluating the features of a post and their relevance to a document class with the aim of identifying natural groupings. A *Fuzzy SVM* model was then used to classify each post using the membership matrix generated by *K-FCM*. This design was aimed at eliminating the inherent noise in social media data, thus improving the accuracy of the detection process. In another experiment, they adopted a semi-supervised learning approach that supplemented an initial training sample with additional training data extracted from unlabelled data. A linear compression voting function was then used to combine the outputs of *Naïve Bayes* and *Stochastic Gradient Descent* classifiers to decide if a post is bullying or not and to enlarge the training set with the labelled output from the classifiers.

Like Nahar *et al.* (2014), Sood *et al.* (2012a; 2012b), and Mangaonkar *et al.* (2015) also introduced voting functions to determine the optimal configuration for cyberbullying detection. Sood *et al.* (2012a; b) developed three profanity detection systems based on three separate features, namely a profanity dictionary, Levenshtein Edit Distance, and Bag-of-words. The profanity dictionary was based on a user-compiled list on phorum.com[7] and noswearing.com. The second system used this profanity list in addition to an edit distance calculator to correct for misspellings. To eliminate false positives, the system checks the words against an English dictionary and a list of names. For example, an edit distance calculator will match 'shirt' to the profane term 'shit' and flag 'shirt' as an offensive term but, by consulting the dictionary, the system will identify the word 'shirt' as not being profanity. The third detection system is an *SVM* classifier that uses bigrams and word stems as features. Running a series of experiments using the three detection systems in various permutations, they obtained their best overall results using a configuration that combines the output of all three systems in an "OR" operation – i.e., if a comment is flagged as profanity by any of the three systems – and the most precise combination used the SVM-based system "AND" either the profanity list or the Levenshtein distance-based system.

Mangaonkar *et al.* (2015) combined the output of multiple classifiers using AND and OR parallelism. They classified tweets using a system consisting of four detection nodes and experimented with homogenous (all computing nodes using the same classification algorithm), heterogeneous (each node uses a different algorithm), and selective (the best performing node is selected as the expert and all other nodes defer to it) collaborations. Each tweet is processed by all nodes and is classified as cyberbullying if more than half of the nodes flag it as bullying in the AND configuration, or if any node flags it as bullying in the OR configuration. They found OR parallelism produces the best *recall* values while AND parallelism provided better *accuracy*. A key aim of their approach is to improve detection speed to facilitate real-time detection using a collaborative computing paradigm over the sequential paradigm more common in cyberbullying detection systems.

The Levenshtein Distance was also used by Nandhini and Sheeba (2015a;b). In (2015b) they combined this with a *Naïve Bayes* classifier to detect cyberbullying on a corpus containing posts from MySpace and spring.me. In (2015a) they substituted Levenshtein Distance with a genetic algorithm to categorise the type of bullying contained in the posts – i.e., flaming, harassing, racism, or terrorism.

Del Bosque and Garza (2014) expressed the aggressiveness of a cyberbullying document as a score, and experimented with lexicon-based, fuzzy systems, and supervised learning detection approaches. For their supervised learning approach, they utilised a *multilayer perceptron neural network* and linear regression and used document length, number of offensive words, and the number of times "*you*" is used as features. They extracted tweets containing specific keywords such as "*school*" from Twitter and then filtered the extracted tweets to those where a user is explicitly referenced via the "@" directive (e.g., "*@username is an idiot*"). They found the best cyberbullying detection was achieved via *linear regression* using document length and offensive words frequency as features. Chavan and Shylaja (2015) also outputted a score representing the probability of a comment being offensive to other users. They used a dataset sourced from Kaggle[8], and a selection of features including skip-grams and combining the results of *SVM* and *Logistic Regression* classifiers.

For their cyberbullying detection system, Zhao and Mao (2016) experimented with an *SDA* (Stacked Denoising Autoencoders) (Vincent *et al.*, 2010) variant called *Semantic-enhanced Marginalized Stacked Denoising Autoencoders* (smSDA) and what they termed *Embedding enhanced Bag-of-Words (EBoW)* (Zhao *et al.*, 2016). They created an initial list of insulting words and used word embeddings to retrieve, from the corpus, words that are most similar to the insulting words. Their approach allowed a *Linear SVM* classifier to learn additional textual features that would otherwise have been deemed of little relevance. For example, the term "*paki*" in the phrase "*be a good paki and say hello*" is an ethnic slur but one that may not be selected as a feature if it is sparsely used within the corpus; if, however, "*paki*" co-occurs with other known cyberbullying words somewhere else within the corpus – for example, in a phrase such as "*you are nothing but a f**king paki*" – then this co-occurrence with a known profane word (i.e., "*f**king*") is used to promote "*paki*" to relevance as a feature. A system such as this can benefit from Parime and Suri's (2014) proposal for a dynamically-sourced profane wordlist that is regularly updated from online resources to

[7] www.phorum.org/phorum5/read.php?16,114701

[8] www.kaggle.com

ensure that new offensive words are captured as they are coined.

Squicciarini *et al.* (2015) used personal, social network and content-specific features with a *C4.5 Decision Tree* classifier to detect bullies on online social networks such as MySpace and spring.me, and devised a set of rules to determine if a user's cyberbullying behaviour is instigated by the actions of another bully. Similarly, Huang *et al.* (2014) found that including social features mined from a user's ego networks as input features to *J48*, *Naïve Bayes*, *SVM,* and *ZeroR* classifiers improved cyberbullying detection over the use of textual features alone. To detect bullying content in their sample of 1000 emails, Burn-Thorton and Burman (2012) found, however, that clustering using a k*NN* algorithm was sufficient.

Hosseinmardi *et al.* (2015) proposed a cloud-based architecture for a scalable detection system for a large social network platform like Instagram. They used n-grams as input features to an *SVM* classifier and network-based features such as "number of followers", "number of followings", and "number of likes" alongside image features to a *Naïve Bayes* classifier, and found the *Naïve Bayes* classifier to be four times faster in predicting cyberbullying instances that the *SVM*. Rafiq *et al.* (2015) also used a *Naïve Bayes* classifier along with *AdaBoost*, *Decision Tree,* and *RandomForest* classifiers to detect cyberbullying instances in Vine; they achieved a 76.39% *accuracy* with *AdaBoost* using unigrams, comments, profile and media information as features.

Potha and Maragoudakis (2014) is one of the few examples in our sample where time was taken into consideration when detecting cyberbullying. They modelled the data as a time series of remarks directed by predators to victims at different points in time. Their dataset consisted of transcripts of online conversations between sexual predators and victims obtained from Perverted-Justice[9], a non-Government organisation that investigates and exposes online sexual predators. They experimented with three feature representation formats namely BoW, weights assignment using SVM, and feature space reduction using Singular Value Decomposition (SVD). They also measured the similarity between conversations held at different times by applying Dynamic Time Warping (DTW) to compute the distance between the time series. *Multi-Layer Perceptron (MLP)* neural network and *SVM* with linear and polynomial kernels were used for predicting cyberbullying instances and *MLP* was found to provide better predictions across all representation formats.

Galán-García *et al.* (2014) focussed their paper on identifying Internet trolls on Twitter using authorship identification techniques. Their approach is based on the hypothesis that an online bully will often create a 'fake' profile specifically for harassing other users, and that the fake profile will be linked to the 'real' profile of the bully. Using the case study of a school where a Twitter account was being used to troll several students, they retrieved all the tweets posted by the trolling account and its followers (17,536 tweets from 92 users) and used four classifiers (*SVM-PolyKernel, J48, SVM-NormalizedPolyKernel,* and *RandomForest*) to analyse the data using the following features: tweets; time of posting; language; geo-position; and Twitter client used. A student from the three users who were consistently ranked as likely authors of the offending tweets by all four classifiers was later revealed as the Internet Troll with the help of the school's authorities. While authorship identification techniques offer potential means of identifying cyberbullies, the study's key assumption that a trolling account will be linked to the real account of the Internet troll will not always be true (a fact acknowledged by the study authors); nevertheless, the incorporation of language forensic techniques into cyberbullying detection systems is an area worthy of further research.

Nahar *et al.* (2012) included sentiment features generated by applying *Probabilistic Latent Semantic Analysis (PLSA)* (Hofmann, 1999) to bullying posts alongside BoW features to train a *Linear SVM* classifier. They found cyberbullying detection improved with the inclusion of sentiment features compared to when only BoW features were used. Nahar *et al.* (2013) achieved even better results by substituting a weighted TFIDF scheme for the bag-of-words (BoW) feature and used Latent Dirichlet Allocation (LDA) (Blei *et al.*, 2003) instead of PLSA to identify sentiment features. Sanchez and Kumar (2011) used a *Naïve Bayes* classifier on tweets extracted by querying Twitter for homophobic slurs and then detected tweets with negative polarity. While such techniques have been successfully used to detect cyberbullying instances, they are rarely sufficient on their own to accurately and consistently identify bullying episodes.

Munezero *et al.* (2014) theorised that including sentiment-based features would improve the detection of anti-social documents. Thus, they expanded on their earlier work (Munezero *et al.*, 2013) by introducing emotion-based features to three classifiers, namely *Naïve Bayes*, *SVM*, and *J48* classifiers. The effect of the inclusion of these features was, however, marginal compared to earlier experiments performed without sentiment-based features (Munezero *et al.*, 2013). This inability of isolated sentiment analysis techniques to accurately detect cyberbullying can be inferred from the work of Xu *et al.* (2012a). They trained four text classifiers (*Naïve Bayes, SVM (linear), SVM (RBF)* and *Logistic Regression*) on a Twitter corpus to identify bullying tweets and the roles played by people referenced within the tweets. By reviewing a subset of the extracted tweets, they detected seven emotions in the tweets, namely anger, embarrassment, empathy, fear, pride, relief, and sadness, and found that, while fear is the emotion most expressed in the tweets (Xu *et al.*, 2012b), it is often jokingly expressed. It would appear from our review that, when used in isolation for cyberbullying detection, sentiment analysis techniques struggle to distinguish between genuine emotions and those sarcastically expressed in bullying messages. We found that mixed-initiative approaches (discussed in a later section) provide a way to improve sentiment-based (and other) cyberbully detection approaches by injecting human-based logic into the detection process.

[9] Perverted-justice.com

In summary, among supervised learning approaches, the most commonly used classifier is SVM. Other frequently used models include *Naïve Bayes* and decision trees such as *J48*. The high utilisation of both *SVM* and *Naïve Bayes* in our sample reinforces their popularity in text classification tasks. Both models are sensitive to parameter optimisation and will outperform one another under different conditions, including features used, the percentage of missing data, and computational speed. *Naïve Bayes* is often desired for its high speed, although its core assumption of independence of attributes may be found too constraining in certain situations. While *SVM* may be preferred when working in high-dimensional spaces, they are often inefficient to train. Tree Ensembles, like *J48*, can handle non-linear features, identify statistical relationships between input and outputs, and compute the strength of such relationships; without additional engineering, however, they will often prune out low occurring instances from within a sample. We did not find a single machine learning method which consistently outperforms the others. Almost all of the supervised learning approaches require careful feature engineering. Recent advances in deep learning have, however, made it possible to train effective classifiers without expensive feature engineering, and this could be a potential future research direction in cyberbullying detection.

2.3.3.2 Lexicon-Based Approaches

Pérez *et al.* (2012) and Fahrnberger *et al.* (2014) developed IM (Instant Messaging)-based cyberbullying prevention tools centred on lexicon-based approaches. MISAAC (Pérez *et al*., 2012) is a multi-agent system designed to detect bullying content in instant messages. The messages are initially processed by a lexical analyser and compared against established patterns of aggression in a content-analysing module. A simple traffic light system is then used to designate the author as green, yellow, or red, indicating an ascending range of aggressiveness based on the message's content. After each message, the sender's colour is recalculated. For example, a green user with no history of aggressiveness sending a mildly offensive message for the first time will still retain his/her green status, but if another offensive message was sent by the same user, the green status will be changed to yellow and further violations will transit the user to a red status. Restrictions are then applied to users based on their current status, as signified by the assigned colour.

Fahrnberger *et al's.* (2014) SafeChat focuses on validating a user's identity. It uses the SecureString 2.0 cryptographic system (Fahrnberger, 2014) to encrypt messages before sending them to the recipient. The system uses a blacklist sourced from WordNet and substitutes bad words detected within the message with safer alternatives. The sender's identity is verified against information held about the user in the system, and the message is only transmitted after the sender's identity has been verified and it has been ascertained that the sender is authorised to contact the recipient. Kontostathis *et al.* (2013) compiled a list of "bad words" based on noswearing.com and queried formspring.me[10] for posts containing any of the words on the list. Essential Dimensions of LSI (EDLSI), a vector space extension (Kontostathis and Pottenger, 2006), was then applied to the corpus to extract the meaning of phrases based on word co-occurrence, and posts were scored based on the number of bad words they contain and their relationship to rest of the message.

2.3.3.3 Rule-Based Approaches

Serra and Venter (2011) proposed a detection system to identify children at risk of cyberbullying by feeding a neural network their mobile phone usage patterns and interpreting set rules linking phone usage patterns to cyberbullying activities. Chen *et al.* (2012) incorporated features such as conversation history and writing style in the development of their Lexical Syntactic Framework (LSF). Using lexical and syntactic features, the system computes an offensiveness score for each sentence within a YouTube comment using a set of rules. A similar score is generated for the user as well, based on the user's writing style and past comments. The syntactic features were mined using the Stanford-parser[11] to extract word-pairs and their type dependencies (e.g., the number of pronouns and offensive words) from each sentence. The dataset consists of comments made by over 2 million distinct users on the top 18 videos across a number of categories. On comparison to *Naïve Bayes* and *SVM* classifiers, experimental results indicate that LSF outperforms traditional machine learning approaches in terms of *precision*, *recall,* and F_1.

Mahmud *et al.* (2008, pg.1) posited that "insulting or abusive messages are an extreme subset of subjective language" and that, by interpreting the basic meaning of a sentence, it is possible to distinguish between expressions containing abusive content and informative text. Like Chen *et al.* (2012), they used the Stanford-parser to detect dependencies between words and used rules based on the discovered dependencies to classify a sentence as abusive or not. Bretschneider *et al.* (2014) also formulated rules to recognise word patterns that indicate a relationship between profane words and person references. They extracted tweets from Twitter and used a Person Identification Module on each tweet to discover any person references via usage of personal pronouns (e.g., "*f**k you*"), name (e.g., "*Tom is an idiot*"), the author's point of view (e.g., "*my chemistry teacher is an a***"), and direct reference to a user (e.g., "*@user, just die!*"). A tweet is then classified as cyberbullying if it matches one or more patterns. When compared with baseline wordlist- and machine learning-based systems, the pattern-based system achieved a 0.15 and 0.09 improvement in the respective F_1 values.

2.3.3.4 Mixed-Initiative Approaches

Following on from their 2011 work, Dinakar *et al.* (2012) attempted to detect indirect bullying messages by incorporating common sense reasoning into their detection system. The common sense reasoning was implemented as a set of over 200 assertions converted into a sparse matrix

[10] www.spring.me

[11] nlp.stanford.edu/software/lex-parser.shtml

representation of concepts versus relations (referred to as their BullySpace Knowledgebase). For each document in the dataset, a set of concepts were extracted and compared to the canonical concepts represented in the BullySpace Knowledgebase. Thus, a message such as "*did you go lipstick shopping with your mum today*" sent to a heterosexual male will be matched to the assertion "*lipstick is used by girls*" and then flagged as an instance of implicit cyberbullying indicative of homophobic sentiments. This method is an example of a mixed-initiative approach to cyberbullying detection, allowing the inclusion of human-based reasoning within the detection process. A bullying message such as this will normally go undetected in many traditional cyberbullying detection systems as it contains neither profanity nor negative sentiments. While this method is heavily reliant on the human knowledge contained within its knowledge base, it certainly offers an avenue to improve traditional detection methods by incorporating real-world human knowledge.

Dadvar *et al.* (2014) also adopted a mixed-initiative approach to cyberbullying detection by using a panel of cyberbullying experts to provide weightings to a feature-set of user-based information such as the age of the user, membership duration, the number of uploads, the number of subscriptions, the total number of posts, and length of the post. The human experts rated each feature on its relative importance and the likelihood that a bully can be identified by the feature. This information was then provided to an expert system called the *Multi-Criterion Evaluation System* (MCES). MCES combines multiple information sources for decision-making support and was used to compute a "*bulliness*" score for each user in a YouTube sample using the weighted scores of the ratings provided by the expert panel for each feature. These features, along with the content-based features used in Dadvar *et al.* (2013a), were then used to train *Naïve Bayes*, *C4.5 decision trees,* and *SVM (Linear)* classifiers. The outputs of MCES and the classifiers were then combined in a hybrid system and they discovered that, while the hybrid system did achieve better performance over both MCES and the ML (Machine Learning) classifiers, it was only marginally better than the expert system.

Finally, Sheeba and Vivekanandan (2013) proposed supplementing their cyberbullying detection system with human knowledge sourced from cyberbullying experts. Using a *Maximum Entropy* classifier, they extracted keywords from a document and then used the extracted keywords to determine the sentiment expressed by the document. A fuzzy system is then used to apply rules created by human experts to make a bullying decision on a document. A topic detection module will then identify each message's topic using calculated word frequencies.

2.3.3.5 Other Approaches

Aside from supervised (and semi-supervised) learning, mixed-initiative, rule-based, and lexicon-based cyberbullying detection systems, we found a number of papers that employ approaches that do not easily lend themselves to our categorisations. Such papers include Bosse and Stam (2011) where the authors formulated the detection problem as a norm violation issue by introducing a number of normative agents into a virtual environment to monitor the activities of users within the virtual world. The agents monitor and record all activities within the environment and use a rewards and punishments system to enforce the desired behaviour for all users. Each user is assigned a reputation score based on their observed behaviours. Norm-violating behaviours such as bullying or stalking negatively affect a user's reputation score and result in punitive actions against the user. The study used a small, purpose-built virtual world and it remains to be seen if the same experiment can be successfully replicated on large online social networks with millions of users like Facebook and MySpace. In addition, for such large virtual environments, the range of actions available to the agents will be limited to what is programmatically possible via the API (Application Programming Interface) exposed by the service providers.

Mancilla-Caceres *et al.* (2012; 2015) also studied user interactions within a virtual environment. They created a social computer game that required players to create teams and work collaboratively together to perform tasks. Using 5th-grade students as case studies, they observed the students' behaviours within the game and compared this to the results of a survey administered by cyberbullying experts to the same group of students prior to the game. By analysing interactions within the game, they discovered a collective attempt by a number of students to bully another student. Interestingly, none of the bullies were flagged by the cyberbullying experts as exhibiting bullying tendencies from the analysis of the survey responses. While such interactions within games and virtual worlds as studied by Mancilla-Caceres *et al.* (2012; 2015) and Bosse and Stam (2011) offer an interesting insight into cyberbullying behaviour, care should, however, be taken when interpreting such data because certain seemingly inappropriate behaviour may be normal within a game-playing context. For example, within multi-player gaming worlds such as *Call of Duty* and *World of Warcraft*, players will often ridicule opposing players (referred to as "*trash talk*") in an attempt to force an error.

Kwak *et al.* (2015) define such anti-social game-playing behaviour as toxic playing and, whilst such behaviours are generally unwelcome and may indeed share certain elements with cyberbullying (for example, a deliberate intent to cause offence), they may be an accepted feature of game-playing communities. Cyberbullying should, therefore, be considered relative to the context and environment where it occurred.

TABLE 3: STUDIES AND DATASETS USED.

Study	Corpus Type	Data Source	Data Size	Annotation Judgement	Dataset Availability
Mahmud et al., 2008	N/A	N/A	N/A	Researchers	N/A
Yin et al., 2009	Websites, blogs and forums	Slashdot, Kongregate, Myspace	4802	Researchers	N/A
Bosse and Stam, 2011	Computer Games	Child Time Machine		Researchers	N/A
Dinakar et al., 2011	Media platforms	YouTube	4500	Experts, Annotators	N/A
Sanchez and Kumar, 2011	Social network	Twitter	5000+	Researchers, MTurk	N/A
Serra and Venter, 2011	Emails, SMS and chat	Phone records			N/A
Burn-Thorton and Burman, 2012	Emails, SMS and chat	Emails	1000+	Researchers	N/A
Chen et al., 2012	Media sharing	YouTube	1700	N/A	N/A
Dadvar and De Jong, 2012	Social network	Myspace	381000	Annotators	http://caw2.barcelonamedia.org/?page_id=98
Dadvar et al., 2012a	Social network	Myspace	381000	Annotators	http://caw2.barcelonamedia.org/?page_id=98
Dinakar et al., 2012a	Social network, Media platforms	spring.me, YouTube	4500+	Experts, Annotators	N/A
Mancilla-Caceres et al., 2012	Computer Games	Computer Game		Annotators	
Nahar et al., 2012	Social network, Websites, blogs and forums	Slashdot, Kongregate, Myspace	575	Researchers	http://caw2.barcelonamedia.org/?page_id=98
Perez et al., 2012	N/A	N/A	N/A	N/A	N/A
Sood and Churchill, 2012a	Websites, blogs and forums	Yahoo!Buzz	1655131	MTurk, Researchers	N/A
Sood and Churchill, 2012b	Websites, blogs and forums	Yahoo!Buzz	1655131	MTurk	N/A
Xu et al., 2012a	Social network	Twitter	1762	Annotators	http://research.cs.wisc.edu/bullying/data.html
Xu et al., 2012b	Social network	Twitter, Wikipedia	3001427	N/A	http://research.cs.wisc.edu/bullying/data.html
Dadvar et al., 2013a	Media platforms	YouTube	54050	Annotators	N/A
Dadvar et al., 2013b	Media platforms	YouTube	4626		N/A
Kontostathis, 2013	Social network	spring.me	24134	MTurk	N/A
Munezero, 2013	Websites, blogs and forums	ISEAR, Wikipedia, movie reviews, Antisocial behaviour (ASB) corpus	803	Researchers	http://www.affective-sciences.org/researchmaterial http://www.cs.cornell.edu/people/pabo/movie-review-data/
Nahar et al., 2013	Social network, Websites, blogs and forums	Slashdot, Kongregate, Myspace	N/A	Researchers	http://caw2.barcelonamedia.org/?page_id=98
Sheeba and Vivekanandan, 2013	Social network	Instant Messaging, blog, Twitter, Facebook	N/A	N/A	N/A
Bretschneider et al., 2014	Social network	Twitter	793	Annotators	http://www.ub-web.de/research/index.html
Dadvar et al., 2014	Media platforms	YouTube	54050	Annotators	N/A
Del Bosque and Garza, 2014	Social network	Twitter	111, 381	Annotators	N/A
Fahrnberger et al, 2014	N/A	N/A	N/A	N/A	N/A
Huang et al., 2014	Social network	Twitter	900,000	Annotators	http://caw2.barcelonamedia.org/?page_id=98
Nahar et al., 2014	Websites, blogs and forums, Social network	Slashdot, Kongregate, Myspace			http://caw2.barcelonamedia.org/?page_id=98

Munezero, 2014	Websites, blogs and forums	ISEAR, Wikipedia, movie reviews, Antisocial behaviour (ASB) corpus	803	Researchers	http://www.affective-sciences.org/researchmaterial http://www.cs.cornell.edu/people/pabo/movie-review-data/
Parime and Suri, 2014	Social network	Myspace	N/A	N/A	N/A
Potha and Maragoudakis, 2014	Websites, blogs and forums	Perverted-Justice	N/A	Annotators	N/A
Chavan and Shylaja, 2015	Websites, blogs and forums	Kaggle	4000	N/A	N/A
Galán-García et al., 2014	Social network	Twitter	1900	N/A	N/A
Hosseinmardi et al, 2015	Social network	Instagram, ask.fm	49000	CrowdFlower	N/A
Mancilla-Caceres et al., 2015	Computer Game	Computer Game	N/A	Annotators	N/A
Mangaonkar et al., 2015	Social network	Twitter	N/A	N/A	N/A
NaliniPriya and Asswini, 2015		Unknown	N/A	N/A	N/A
Nandhini and Sheeba, 2015a	Social network	Myspace, Spring.me	N/A	N/A	http://caw2.barcelonamedia.org/?page_id=98
Nandhini and Sheeba, 2015b	Social network	Myspace, Spring.me	N/A	N/A	http://caw2.barcelonamedia.org/?page_id=98
Rafiq et al, 2015	Social network	Vine	436000	CrowdFlower, Expert	N/A
Squicciarini et al, 2015	Social network	Myspace, Spring.me	3032	Annotators	N/A
Zhao and Mao, 2016	Social network	Twitter, Myspace	1539	Annotators	N/A
Zhao et al., 2016	Social network	Twitter	1762	N/A	N/A

TABLE 4: CORPUS TYPE AND STUDIES WITH HIGHEST ACCURACY, PRECISION, RECALL AND F-MEASURE SCORES PER DETECTION TASK FOR THE CORPUS TYPE.

Corpus Type					Dataset	Task Performed	Study	Dataset URL
	Accuracy	Precision	Recall	F_1				
Social networks	0.76	N/A	N/A	0.94	spring.me	Binary Classification	Nandhini and Sheeba, 2015b.	http://caw2.barcelonamedia.org/?page_id=98
Websites, blogs and forums	0.996	0.996	0.996	0.996	ISEAR, Wikipedia, movie reviews, Antisocial behaviour (ASB) corpus	Binary Classification	Munezero et. al., 2014	http://www.affective-sciences.org/researchmaterial http://www.cs.cornell.edu/people/pabo/movie-review-data/
Websites, and blogs and forums	0.992	N/A	N/A	N/A	Slashdot, Kongregate, Myspace	Role Identification	Nahar et al., 2012	http://caw2.barcelonamedia.org/?page_id=98
Media platforms	N/A	0.9824	0.9434	0.95	YouTube	Role Identification	Chen *et al.,* 2012	N/A

**N/A – Not Available*

3 Discussion

Our survey covers research efforts on automatic detection of cyberbullying. The review includes articles, published over the last 8 years, starting with the pioneering work of Mahmud *et al.* (2008). The breadth of the studies included in the survey emphasises the growing attention that cyberbullying prevention has been receiving in recent years. Although supervised learning approaches dominate the methods considered by many studies, researchers have demonstrated willingness to utilise emerging work from other areas of natural language processing in order to improve performance. In what follows, we discuss the previous studies from five perspectives, their interpretations of cyberbullying, features used for cyberbullying detection, performance comparison, dataset creation, and preventive actions against cyberbullying.

3.1 Interpretations of Cyberbullying

We found that, while studies generally agree that cyberbullying is an intentional malicious act, there exist slight variations in how researchers interpret cyberbullying for detection purposes. For example, if an act has to be repeated before being considered cyberbullying, then the detection system must maintain a history of previous messages and perhaps introduce the timestamps of messages exchanged as a feature to satisfy the "repeated acts" criterion. Potha and Maragoudakis (2014) and NaliniPriya and Asswini (2015) were the only studies in our sample that incorporated time as part of the detection process. Potha and Maragoudakis (2014) used messages' timestamps as features while NaliniPriya and Asswini (2015) computed changes in a user's social network over a period of time. We envisage that the use of timestamps as features will increase in cyberbullying detection research, especially as time information is easily accessible in all forms of electronic communication. For example, a rule can be created to only flag a user as a cyberbully if he or she exceeds a threshold of bullying messages over a set period of time.

Intent and power differential are two key components of bullying that have proven difficult for researchers to demonstrate within an electronic context. Prior to the determination of these two components, however, is the identification of the victim. Consider the following tweet; "*Going to Africa, hope I don't get AIDS. Just kidding. I'm white!*". While the offensive nature of the tweet is not in dispute, its classification as bullying content is more subjective as the message does not appear to be directed at any particular person, rather its intention appears to be causing offence to an entire continent and race. Establishing that intended victim(s) are distinct entities is, therefore, an important part of cyberbully detection (e.g., tweets addressed to a particular person can be easily extracted via the @username tag).

Once the intended recipient has been successfully identified, the task of classifying an online interaction as cyberbullying necessitates ascertaining that the recipient is indeed a victim and one that cannot easily defend him or herself. This requires establishing a power differential between bully and victim(s). We found no such attempts in our sample. This, in itself, is not surprising as it is a non-trivial task. Understanding the nature of the relationship that exists between the parties involved in a cyberbullying episode can help determine if a power differential does indeed exist. Ego networks and network-based mutual reinforcement algorithms (such as Hyperlink-Induced Topic Search (HITS)) (Kleinberg, 1999) are some of the approaches used by studies in our survey to model relationships within an online social network. These are then typically used to identify bullies and victims within the network based on the frequency and offensiveness of messages exchanged. Such approaches automatically label senders of offensive messages as bullies and the recipients as victims. Alleged victims can, however, reciprocate with equally or more offensive messages. Thus, by computing an offensiveness score for messages sent in both directions it may be possible to judge a victim's ability to defend him or herself and establish if a power differential does indeed exist between the two parties.

This aforementioned approach, however, fails in the assumption that, by responding with offensive material, victims are capable of defending themselves. If a victim does not reply in kind with offensive content does that mean a power differential then exists? Equally, the presence of an offensive reply does not negate the need to establish a power differential. Take, for example, a tweet sent by US presidential candidate Donald Trump ahead of the 2016 US Presidential Elections: "*Sad sack @JebBush has just done another ad on me, with special interest money, saying I won't beat Hillary - I WILL. But he can't beat me*". The tweet calls Jeb Bush, another Presidential candidate, a "*sad sack*". If Jeb Bush chooses to ignore this tweet or reply with a non-offensive tweet, this does not make him any less powerful in this context. Both parties are powerful political figures and there is no obvious power differential in this situation. Hence, how to establish a power differential effectively from electronic exchanges remains an open problem.

Although our inclusion criteria included identifying other cyberbullying roles such as defenders, instigators and bystanders, none of the surveyed research attempted to identify these additional roles within their various experiments. We also did not find studies exploring advanced concepts such as multiple bullies 'ganging up' on a victim, individuals performing multiple roles, or even transitioning between roles. Of the 4 key detection tasks defined for our survey, binary classification occurs more frequently than any other tasks. This is understandable as it is often the first task performed within the process, providing the foundation upon which additional tasks are launched. The proper execution of this task, therefore, takes on additional importance as inaccurate results can corrupt the output of subsequent tasks. For example, a role classifier may wrongly label an individual a bully if the preceding binary classification phase incorrectly flags innocent messages as bullying.

The complicated nature of cyberbullying may, therefore, necessitate the combination of multiple tasks to cap-

ture various forms of cyberbullying. In this regard, the system can begin by performing binary classification using content- and sentiment-based features to identify individual bullying messages. Once a bullying message has been identified, it needs to be established as part of a sequence to satisfy the repeatability criteria and then user- and network-based features can be used to perform role identification and ascertain if a bully-victim power differential does indeed exist. Finally, after the incident has been classified, the online behaviour of the involved parties can be tracked and mapped to their identified roles and used to trigger further intervention if required.

3.2 Features Used for Cyberbullying Detection

An inherent difficulty in detecting cyberbullying is its highly subjective nature. The same message can have different effects on separate individuals and it is very difficult to determine what these effects will be at the time of detection. Approaches such as that of Dadvar *et al.* (2012b) which analyses the victim's follow-on action after receiving a bullying message, as well as research into determining a user's emotional state based on their posting behaviour in social media (De Choudhury *et al.*, 2013), can help in this regard and are certainly worthy of further research.

Content-based features such as spelling, presence of pronouns and profanity, document length, and capitalisation featured heavily in studies such as Dadvar *et al.* (2013a; 2013b; 2014), Dinakar et al. (2012a) and Nahar *et al.* (2014). The usefulness of these features is largely dependent on the corpus and detection task. For example, on a Twitter dataset, the usefulness of document length as a feature will be limited as documents within the dataset will exhibit little variety in length (due to the maximum character limit imposed by Twitter). Likewise, the usefulness of interpreting capitalised words as the textual equivalent of shouting is reliant on frequent occurrences within a dataset. In fact, its occurrence is low within the publicly-available datasets from our sample, and this is likely to be the case for many other datasets as well.

Using pronouns and/or profanity appears to be quite popular in cyberbullying detection research, as evidenced by its use in 25 papers within our sample. This popularity is due to their effectiveness in identifying abusive and insulting content. Existing methods for identifying profanity can, however, be improved by substituting the static wordlists typically used with a dynamic system capable of querying online resources whenever a new term is encountered, thereby ensuring that the profanity list does not become outdated. The use of pronouns and profanity as features alone does not, however, guarantee that all instances detected are cyberbullying; as previously discussed, abusive content is more likely to represent cyberaggression than cyberbullying. Of the 25 papers that use them as features, only Kontostathis (2013) and Bretschneider *et al.* (2014) did not combine them with other features. When combined with other features, their ability to detect cyberaggression (a key component of cyberbullying) makes pronouns and profanity two of the most useful content-based features for cyberbullying detection.

Sentiment-based features provide an exciting avenue to incorporate recent advances in sentiment analysis into cyberbullying detection but more research is required to fully gain the benefits of these features. While the polarity of sentiments expressed in a product review is a good indication of a writer's overall opinion about the product, this is not always the case with cyberbullying. Negative sentiments can be expressed in support of a victim and against a bully (for example, speaking out against racism) and vice versa. Thus, sentiment polarity is only of value if additional context, such as the object of the expressed sentiment and its relationship to the victim, is available. This can be seen in Dinakar *et al.* (2011), where messages were first clustered based on topics (i.e., the object) before using sentiment-features to determine the polarity expressed about these topics. Simply detecting emotions and polarity cannot, therefore, be relied upon to accurately detect cyberbullying.

3.3 Performance Comparison

Classifiers are typically evaluated based on key metrics such as *Accuracy*, *Precision*, *Recall* and F_1. While a number of papers within our sample provided values for these metrics in their experiments, a direct comparison of the studies based on these results is, however, difficult. This is because the datasets used by the studies will have a direct impact on the results. Without conducting the experiments on the exact same dataset, a comparison of the achieved metrics' values is meaningless. Even studies that used the same dataset tend to sample different extracts from within the dataset. We grouped the datasets used by the studies in our survey into 4 categories (see Table 3), namely "*social networks*", "*websites, blogs and forums*", "*media platforms*", and "*email, SMS and chat*".

We found that some of the highest scores achieved were by studies using datasets that fall within the websites, blogs, and forum category. These corpora may not be representative of cyberbullying and, as such, the scores achieved using such corpora cannot be directly compared against those achieved using a more representative sample, such as those in the *social networks* and *media platforms* categories. For example, Nahar *et al.* (2013) achieved F_1 of 0.92 using an *SVM* classifier on a Kongregate dataset while Dadvar *et al.*'s experiments using SVM on MySpace (2012a) and YouTube (2013b) corpora yielded F_1 of 0.28 and 0.64, respectively. Kongregate is a website devoted to video games with a likely low occurrence of cyberbullying, while MySpace and YouTube are social network/media platforms where cyberbullying is likely to be more prevalent.

Rather than comparing the raw values for scores published by each study, we compare the study with the highest F_1 value for each dataset category per each detection tasks (e.g., binary classification and role identification are separate tasks that can be performed on the same datasets, each resulting in different scores) and present (where available) the *Accuracy*, *Precision* and *Recall* scores for these. If the F_1 value is not available for the study, we use the highest value of *Accuracy*, *Precision* and *Recall* in descending order of importance. This information is presented in Table 4. Researchers can use these values as a guide when

conducting experiments using comparable datasets. Thus the best results achieved for binary classification using a social network corpus was by Nandhini and Sheeba (2015b). Likewise, Chen *et al.* (2012) achieved the best role identification scores for media platform type corpora.

3.4 Dataset Creation

Our survey revealed that the majority of datasets used by studies in our sample are more likely to represent online harassment/insults than cyberbullying. These datasets typically contain individual instances of abusive content and are, therefore, unsuitable for creating features to detect repeated acts of aggression or establishing power differential when developing classifiers. They are, therefore, only useful for detecting cyberaggression, and it can be argued that these studies are essentially detecting just one aspect of cyberbullying. It is disappointing that we found very little evidence of researchers going beyond cyberaggression into more complex tasks, such as establishing power differential and repeatability. A key enabler for performing these tasks, however, is the availability of quality datasets with enough datapoints to enable the extraction of features to support the detection of these criteria. As cyberbullying has been shown to proliferate on social media, researchers are more likely to find representative samples of wholistic cyberbullying (as opposed to only cyberaggression) within social media than other types corpora.

Age and gender are used as features in *social network*- and *media platform*-based datasets more than any other information (the ease of extraction very likely contributing to their popularity as features); these can, however, be easily falsified leading to data corruption unless a scheme to verify this information is implemented. As such, when creating user- and network-based features, data automatically generated by the network – such as number of likes, time of posting, and friends' lists – should be preferred over user-provided data like age, gender, and location. For example, Facebook friends lists can be treated as a labelled sample to validate a user's social network, generated using aforementioned algorithms like Ego network and HITS. Equally, posts, comments, and status updates can be tied to specific users and a user's activities across the platform can be extracted and annotated. Thus, in this way, a dataset containing users and a collection of their online activities can be created and used for tasks such as role identification, assessing if a bully's friends exhibit similar tendencies, and the influence of such friends on their online behaviour. For example, do they post offensive messages following similar activities by friends? Are the themes of these messages similar and are they directed to the same type of users (e.g., based on ethnicity, gender or physical appearances)?

It is encouraging to see studies like Dadvar *et al.* (2013;b; 2014), Hosseinmardi *et al.* (2015), Rafiq *et al.* (2015) and Squicciarini *et al.* (2015) using these types of data as features. With the exception, however, of Squicciarini *et al.* (2015) (who detected events following the occurrence of cyberbullying), these features were only used for binary classification and bully-victim role identification. Nevertheless, we can see a trend of studies increasingly mining user- and network-based features from social media for cyberbullying detection and, ultimately, this will advance research into the detection of all aspects of cyberbullying (e.g., repeatability).

Extracting data from social media is, however, not without its challenges. Privacy and ethics concerns are some of the key issues researchers must adequately mitigate before data can be mined from these platforms. In addition, the features of each platform and the intended cyberbullying detection tasks will determine the suitability of a platform as a data source. Twitter, with its publicly available data, may be easier to mine than other platforms such as MySpace, Instagram, or Facebook. The more personal nature of the latter platforms may, however, provide richer data for profiling users. Comments on YouTube videos are likely to contain abusive content by people unknown to the original poster than, for example, a wall post on Facebook. Facebook and Instagram posts, on the other hand, may provide more indication of the potential emotional state of the poster at the time of posting. In addition to textual content, followers, retweets, and other useful information can be mined for dataset creation. For example, Twitter's retweeting and hashtag features could be harnessed to model the propagation of a specific post or meme as it is virally propagated across the network.

Extracting data from Facebook can present a challenge as only the public versions of user profiles can be extracted, and these will typically contain less information than the full profile. User information such as friend lists (a key data point for modelling user relationships) cannot be extracted without the login details and permission of the profile's owner. Facebook, however, provides its Public Feed API which returns public status updates and wall posts across the network. A large dataset can thus be created comprising posts from a wide variety of users. Equally, Instagram's API provides methods to extract publicly-shared media and the associated comments, likes, and tags. Using these APIs, it is, therefore, possible to extract a large amount of data for dataset creation. Once extracted, personal data such as age, gender, location, usernames, and references to named entities can be removed from the data and generic placeholders substituted to ensure that the data is duly anonymised. A unique identifier can be assigned to each user if a post's sender is to be preserved. There are several existing methods to perform this task and these are more than sufficient.

What makes a good cyberbullying dataset? Based on our survey, we suggest as a starting point a social media extract with a minimum of a few thousand individual posts with at least 10%-20% positive cyberbullying instances. If user-based features are of importance then the posts should be attributable to a sufficient number (n = 10% of posts) of users and each distinct user responsible for at least 10 posts. As cyberbullying is prevalent amongst adolescents (Livingston *et al.*, 2014), researchers should target users aged 13–18 when extracting data to increase the likelihood of capturing positive bullying instances within the extracted sample. By instances, we refer to messages, status updates, wall posts, likes, and comments. This is only provided as a guide to assist researchers new to the

field and not a hard rule. Ultimately, researchers, choice of classifiers, and the intended classification tasks will determine how much data is sufficient.

The sheer volume of data that can be extracted from social media introduces labelling challenges as well. Even with the high cyberbullying percentages reported on social media, the number of actual positive samples within an extract will be comparatively small. It may, therefore, not be feasible to use experts or researchers to label all of this data and crowdsourcing (used on its own or in conjunction with other annotation schemes) may be a practical and cost-effective way to label all or the majority of the data. Kontostathis (2013) and Sood and Churchill (2012a;b) successfully used crowdsourcing to annotate data. Studies like Nowak and Rüger (2010) and Snow *et al*. (2008) compared the use of experts versus crowdsourcing in annotating large datasets and found that many large data labelling tasks can be accurately carried out using crowdsourcing, especially if methods to eliminate unreliable labellers, such as those proposed by Raykar and Yu (2012) and Welinder and Perona (2010), are implemented as well.

3.5 Preventive Actions

Cyberbullying detection is a key stage within the larger issue of cyberbullying prevention. Quite often, the detection approach is influenced by the preventive actions intended. Researchers will design detection systems with preventative actions in mind and these play a crucial part in the actual detection process, often influencing how the detection system is tuned and optimised and thus directly affecting the results achieved. For example, if any detected cyberbullying messages will result in severe punishment for the user (e.g., banning from the network) then the penalty for false positives is significantly higher than in a scenario where the messages were to be simply flagged to the user. In the case of the former, researchers may concentrate on improving the system's *Precision* to ensure that as many of the detected instances are actual bullying in order to reduce the possibility of a user being banned for sending a message that was wrongly labelled as bullying. The vast majority of studies discovered in our survey focused almost entirely on cyberbullying detection without recourse to the preventive actions to be taken once bullying is detected, with notable exceptions being Dinakar *et al.* (2012a) and Bosse and Stam (2011). In Dinakar *et al*. (2012a), the authors discussed reflective user interfaces designed to discourage anti-social behaviour by giving would-be bullies cause to pause and rethink their actions. This can be in the form of action delays, whereby the send button is disabled for a few seconds, or highlighting inappropriate parts of a message. In fact, such preventive measures are already in use and can be seen in mobile apps such as RethinkWords[12] and Bully Free Keyboard[13]. Both apps are virtual keyboards that provide simple but effective ways to educate and discourage cyberbullying from the message sender's perspective by detecting when inappropriate words are used and reacting accordingly. RethinkWords displays a message when offensive words are typed, encouraging the sender to reconsider sending such an abusive message, while Bully Free Keyboard temporarily disables the keyboard for a few seconds after an inappropriate term has been typed, thus serving as a constant reminder of the negative impact of such words. The efforts of these apps could be taken further by incorporating some of the more advanced techniques discussed in this paper, rather than the simple lexicon-based approach that is currently being used. Bosse and Stam's (2011) agent-based system provides another interesting way to encourage positive behaviours in would-be bullies. Their use of BDI agents is unique in our sample, and can certainly be evolved to subtly discourage cyberbullying even without the bullies realising it, especially for pre-teen users.

Finally, while the classification of resulting events after a cyberbullying incident is still relatively novel, it offers tremendous possibilities, especially if married to ongoing works in areas such as the management of digitally manifested distress. This will improve the practical usefulness of these detection systems since detecting cyberbullying is only a part of the process; the full benefits are derived by enabling the appropriate actions to be more easily taken once a match is made. For example, notifying responsible adults or mental health professionals when a cyberbullying victim starts exhibiting tell-tale signs of distress following a cyberbullying attack.

4 Research Challenges

The lack of a universally-adopted cyberbullying definition for detection purposes and a dearth of large, labelled cyberbullying corpora are two key research issues facing cyberbullying detection research.

4.1 Non-Holistic Consideration of Cyberbullying

While the majority of researchers agree on the definition of cyberbullying to include the core criteria of repetitiveness, intent to cause harm, and power differential, we found little evidence of studies tackling cyberbullying in such a holistic manner. Researchers, we found, often equate the detection of any form of abusive and offensive content to the detection of cyberbullying with little or no attempt to establish an intent to cause harm, a power differential, or the repetitive nature of the offensive acts. To progress the state-of-the-art beyond binary classification of abusive content, it is crucial for researchers to embrace the holistic definition of cyberbullying for detection purposes.

4.2 Inadequacy and Lack of Cyberbullying Datasets

The challenge posed by the lack of easily-accessible labelled corpora is underscored by the fact that our survey revealed only 5 distinct publicly-available datasets. Messaging-focused social media platforms like Instagram, Snapchat, and Whatsapp are under-represented in these datasets, and corpora based on these platforms will cer-

[12]https://itunes.apple.com/us/app/rethink-stop-cyberbullying/id1035161775?mt=8

[13]https://itunes.apple.com/gb/app/bully-free-keyboard/id977170220?mt=8

tainly be welcomed by the cyberbullying research community. Such datasets can include entire conversations amongst several users and feature multiple annotation schemes. For example, annotations can be by users and their roles (enabling classifiers to be trained to recognise the different roles in cyberbullying), by bullying type (direct or indirect bullying), by conversations (i.e., labelling the relationship that exists between two people based on messages exchanged), and by timestamp (including the time each message was sent to enable the generation of a time series to satisfy the repetitive nature criterion. The latest version of the Bullying Traces Dataset[14] is a step in this direction as it includes, in addition to the binary labelling of each tweet as bullying or not, the roles played by the tweets' authors, emotions expressed in the tweet, and the type of bullying trace the tweet constitutes (i.e., reportage of cyberbullying, teasing, actual cyberbullying, and accusation of bullying against another user). This is a significant improvement over other datasets as it allows for more advanced detection tasks such as role and bullying type identification. Future attempts at creating cyberbullying datasets should endeavour to emulate efforts such as these.

5 FUTURE DIRECTIONS

Following on from our review of current literature, we recommend some future directions to advance cyberbullying detection research.

5.1 Detection of Non-Textual Cyberbullying

While the focus of the studies in our sample has largely been on textual bullying, images and videos can also be used as delivery systems for online bullying and their impact can be as, or perhaps even more, damaging. In addition, as social media platforms improve their ability to detect and prevent textual bullying, bullies may likely resort to the use of other media forms to bypass anti-bullying measures. Recent advances in image processing and OCR (Optical Character Recognition) make it viable to attempt cyberbullying detection within media forms like images, animations, and videos. With social media trends such as internet memes and viral videos becoming hugely popular in recent times, these can be easily perverted by bullies to perpetrate cyberbullying. We, therefore, envisage that developing systems capable of detecting bullying content within multimedia files is a key area for future research considerations.

5.2 Expanding Cyberbullying Role Detection beyond Victims and Bullies

When cyberbullying occurs, there are often multiple roles at play beyond the traditional roles of bullies and victims. These include roles such as instigators, defenders, and bystanders. A future research direction would be to extend role identification to map these additional roles and track if and how individuals change or adopt additional roles within the course of the bullying episode. For example, do bystanders eventually become bullies or evolve to defenders? How are coordinated attacks involving multiple bullies orchestrated – are they planned and do the bullies communicate prior to an attack? What was the initiating event? Can this event be attributed to a post/comment made by the victim? These are questions as yet unanswered by researchers.

5.3 Determining a Victim's Emotional State after a Cyberbullying Incident

Determining events and a victim's emotional state after a cyberbullying incident is another emerging research area. Research in this area is important because when a bullying instance goes undetected, it is equally important to manage the end results of such situations. Analysing the victim's immediate response to online interactions may unearth clues as to the nature of the initial interaction. For example, a victim may change his/her profile details following such interactions, post content containing negative sentiments, or leave the network abruptly. Such instigating interaction can be flagged up for subsequent review by a human who can then follow-up with appropriate actions. The process can also be used as a feedback loop to manually apply the correct label to the undetected bullying incident for re-training the classifier. This can assist in providing much-needed support for cyberbullying victims.

5.4 Word Representation Learning for Cyberbullying Detection

Recent advances in word representation learning (Mikolov *et al.*, 2013) have made it possible to build text classifiers from word representations or word embeddings trained from large corpora, such as Wikipedia or Google News Corpus. Deep neural networks have proven effective in learning non-linear feature transformations in generating word embeddings. Such word embeddings could be beneficial to cyberbullying detection. Experiments can be performed to generate word embeddings from different datasets, ranging from general corpora (e.g., Wikipedia) to more specialised datasets (e.g., abusive tweets) to compare their effectiveness for cyberbullying detection. Also, while deep neural network approaches require large-scale data for training, traditional methods such as Singular Value Decomposition (SVD) can be explored to learn domain-specific word embeddings from word co-occurrence matrices derived from small-scale data.

5.5 Detecting Cyberbullying in Streaming Data and Real-time

Our survey revealed that the standard approach in cyberbully detection research is to train and evaluate classifiers on static data collected at a point in time. The results published for these experiments give no indication, however, of how well such classifiers will perform in real time in terms of detection speed and ability to cope with streaming data. For example, consider instant messaging platforms such as Instagram and Whatsapp to be effective on such a platform, a cyberbullying detection system must be able to

[14] http://research.cs.wisc.edu/bullying/data.html

classify messages in a timely manner as they are delivered to the user. Using APIs such as the Twitter Streaming API, which broadcasts continuous streams of data as it is generated, a classifier can be evaluated on how quickly it is able to detect cyberbullying events as they occur within the stream. Studies such as Xu *et al.* (2012a) already use Streaming API to source training data, thus using it for evaluation is a natural extension.

5.6 Evaluating Annotation Judgement

Supervised learning was the most popular approach for cyberbullying detection in our survey. The means by which annotation judgement is achieved is of importance when adopting this approach. While research exists that compares and evaluates expert annotation and annotation via crowdsourcing for data mining tasks like image analysis (Nowak and Rüger, 2010), we found no such studies for cyberbully detection. A future research direction can, therefore, be to evaluate and compare quality when performing annotations via these methods, and how the choice of annotation judgement affects detection results.

6 CONCLUSIONS

Popularised by technology companies like Google, Amazon, Microsoft, and start-ups like MetaMind[15], Machine Learning is currently enjoying significant attention in both commercial and research communities. There is a wealth of ML resources available to researchers at little to no cost. For example, a large-scale neural network can be created and deployed within minutes using Microsoft Azure or Amazon Web Services, an unthinkable feat ten years ago. Our survey revealed that there is a growing and impressive body of research on cyberbullying detection, but more work is required to advance the area.

Binary classification of messages as bullying or not and bully-victim identification are the most common tasks performed by researchers. These tasks can be considered well-researched and researchers now need to direct efforts towards more advanced tasks, such as detecting cyberbullying via social exclusion, proving power differential and repeatability criteria, identifying other roles, and mapping actors' transition from one role to another during a cyberbullying episode. We have also discovered that profanity and abusive text is often equated to bullying. This is not always the case and more effort should be directed at detecting cyberbullying in text devoid of profanity and insults.

Compared to the available body of work, the number of publicly-available cyberbullying datasets is low. In addition, many of the publicly-available datasets are outdated and more can be done to ensure researchers entering the field do not go through pains associated with acquiring quality data. It is telling that, while machine learning applications such as facial recognition, personalised movie, and music recommendations are now common features of everyday life, major social media platforms are still reliant on "Report Abuse" buttons to combat cyberbullying.

Cyberbullying is an issue of great importance, one that affects the lives of many young people. The current state of affairs for cyberbullying prevention within online social networks therefore requires urgent attention and improvement. This improvement is only possible if the research community, educational institutions, law enforcement, social media platforms, and software vendors make conscious and concerted efforts to facilitate the diffusion of knowledge and expertise in all directions. It is only when this happens that viable cyberbullying detection applications can advance beyond research boundaries into the wider world.

[15] www.metamind.io

REFERENCES

1. Abeele, M.V. and De Cock, R. (2013). Cyberbullying by Mobile Phone among Adolescents: The Role of Gender and Peer Group Status. *Communications*, 38(1), p.107-118.
2. Al Mazari, A. (2013). Cyber-bullying taxonomies: Definition, forms, consequences and mitigation strategies. IN: International Conference on Computer Science and Information Technology (CSIT). 5th. Amman. March 27 – 28, 2013. IEEE, 126-133.
3. Argamon, S., Koppel, M., Fine, J. and Shimoni, A. R. (2003) Gender, Genre, And Writing Style In Formal Written Texts. *Text Interdisciplinary Journal for the Study of Discourse*, 23, p.321-346.
4. Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3, p.993-1022.
5. Bosse, T. and Stam, S. (2011). A normative agent system to prevent cyberbullying. IN: IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. Lyon, France, August 22-27, 2011. IEEE, 425-430.
6. Bradley, M.M. and Lang, P.J. (1999). Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings. *Technical Report C-1*, The Center for Research in Psychophysiology, University of Florida, p.1-45.
7. Bretschneider, U., Wöhner, T., and Peters, R. (2014). Detecting Online Harassment in Social Networks [online]. Available from http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1003&context=icis2014 [Accessed 20th March 2015]
8. Burn-Thornton, K. and Burman, T., 2012, November. The Use of Data Mining to Indicate Virtual (Email) Bullying. IN: Global Congress on Intelligent Systems (GCIS). 3rd. Wuhan. November 6 – 8, 2012. IEEE, 253-256.
9. Chavan, V.S. and Shylaja, S.S. (2015). Machine learning approach for detection of cyber-aggressive comments by peers on social media network. IN: International Conference on Advances in Computing, Communications and Informatics (ICACCI). Kerala. August 10-13, 2015. IEEE, 2354-2358.
10. Chen, Y., Zhou, Y., Zhu, S., and Xu, H. (2012). Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. IN: International Conference on Privacy, Security, Risk and Trust (PASSAT) and Social Computing (SocialCom). Amsterdam, September 3-5, 2012. New York: IEEE.
11. Cohen, R., Lam, D.Y., Agarwal, N., Cormier, M., Jagdev, J., Jin, T. and Wexler, M. (2014). Using Computer Technology to Address the Problem of Cyberbullying. *ACM SIGCAS Computers and Society*, 44(2), p.52-61.
12. Dadvar, M. and De Jong, F. (2012). Cyberbullying detection: A Step toward a Safer Internet Yard. IN: International conference companion on World Wide Web. 21st. Lyon, April 16 - 20, 2012. London: ACM, 121-126.
13. Dadvar, M., De Jong, F.M.G., Ordelman, R. J. F. and Trieschnigg, R. B. (2012a). Improved Cyberbullying Detection Using Gender Information [online]. Available from http://eprints.eemcs.utwente.nl/21608/01/DIR12_reviewed04.pdf [Accessed 5th November 2014].
14. Dadvar, M., Ordelman, R., De Jong, F. and Trieschnigg, D. (2012b). Towards User Modeling In The Combat Against Cyberbullying. *Natural Language Processing and Information Systems*, p.277-283.

15. Dadvar, M., Trieschnigg, D. and De Jong, F. (2013a). Expert Knowledge for Automatic Detection of Bullies in Social Networks [online]. Available from http://doc.utwente.nl/88358/1/paper_79.pdf [Accessed 5th November 2014].
16. Dadvar, M., Trieschnigg, D., Ordelman, R. and de Jong, F. (2013b.) Improving cyberbullying detection with user context. IN: European Conference on Information Retrieval. 35th. Moscow. March 24th – 27th, 2013. Springer Berlin Heidelberg, 693 - 696.
17. Dadvar, M., Trieschnigg, D., and De Jong, F. (2014). Experts and Machines against Bullies: A Hybrid Approach to Detect Cyberbullies. *Advances in Artificial Intelligence*, 8436, p.275-281.
18. Dadvar, M., Trieschnigg, D., Ordelman, R. and De Jong, F. (2013). Improving Cyberbullying Detection with User Context. *Advances in Information Retrieval*, p.693-696.
19. Del Bosque, L.P. and Garza, S. E. (2014). Aggressive Text Detection for Cyberbullying. *Human-Inspired Computing and Its Applications*, p.221-232.
20. De Choudhury, M., Gamon, M., Counts, S. and Horvitz, E. (2013). Predicting Depression via Social Media [online]. Available from https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/icwsm_13.pdf [Accessed 21st June 2015]
21. Dinakar, K., Jones, B., Havasi, C., Lieberman, H. and Picard, R. (2012a). Common Sense Reasoning For Detection, Prevention, and Mitigation of Cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3).
22. Dinakar, K., Jones, B., Lieberman, H., Picard, R., Rose, C., Thoman, M. and Reichart, R. (2012b). You too?! mixed-initiative lda story matching to help teens in distress. IN: International AAAI Conference on Weblogs and Social Media (ICWSM 2012). 6th. Dublin. June 4 – 7, 2012. AAAI, 74 – 81.
23. Dinakar, K., Reichart, R. and Lieberman, H. (2011). Modeling the Detection Of Textual Cyberbullying. The Social Mobile Web [online] Available from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/3841Karthik/4384 [Accessed 10th February 2015].
24. Ditch The Label (2013) The Annual Cyberbullying Survey [online]. Available from http://www.ditchthelabel.org/downloads/the-annual-cyberbullying-survey-2013.pdf [Accessed 21st June 2015].
25. Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A Publicly Available Lexical Resource for Opinion Mining. IN: Language Resources and Evaluation (LREC). Genoa. May 24-26, 2006. Paris: ELRA, 417-422
26. Fahrnberger, G. (2013). Securestring 2.0-A Cryptosystem For Computing On Encrypted Character Strings In Clouds. *Innovative Internet Community Systems*, 10, p.226 – 240.
27. Fahrnberger, G., Nayak, D., Martha, V. S. and Ramaswamy, S. (2014). SafeChat: A Tool to Shield Children's Communication from Explicit Messages. IN: International Conference on Innovations for Community Services (I4CS). 14th. Reims. June 4 -6, 2014. New York: IEEE, 80 – 86.
28. Fanti, K.A., Demetriou, A.G., and Hawa, V.V. (2012). A Longitudinal Study of Cyberbullying: Examining Risk and Protective Factors. *European Journal of Developmental Psychology*, 9(2), p.168-181.
29. Galán-García, P., de la Puerta, J.G., Gómez, C. L., Santos, I. and Bringas, P.G. (2014). Supervised Machine Learning for the Detection of Troll Profiles in Twitter Social Network: Application to a Real Case of Cyberbullying. IN: International Joint Conference SOCO'13-CISIS'13-ICEUTE'13. Salamanca. September 11 -13, 2014. London: Springer International Publishing, 419-428.
30. Garlan-Garcia, M., Gamon, M., Counts, S. and Horvitz, E. (2013). Predicting Depression via Social Media. (p. 2). IN: International AAAI Conference on Weblogs and Social Media (ICWSM). Boston. July 8-11, 2013
31. Hinduja, S and Patchin, J.W. (2009) Bullying Beyond The Schoolyard: Preventing And Responding To Cyberbullying. Thousand Oaks: Sage.
32. Hinduja, S. and Patchin, J.W. (2008). Cyberbullying: An Exploratory Analysis of Factors Related To Offending and Victimization. *Deviant behavior*, 29(2), p.129-156.
33. Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. IN: Conference on Uncertainty in artificial intelligence. 15th. Stockholm. July 30 – September 1, 1999. San Francisco: Morgan Kaufmann Publishers Inc, 289-296.
34. Honjo, M., Hasegawa, T., Hasegawa, T., Suda, T., Mishima, K. and Yoshida, T. (2011). A framework to identify relationships among students in school bullying using digital communication media. IN: International Conference on Privacy, Security, Risk and Trust (PASSAT) and Social Computing (SocialCom). Boston. 9-11 October, 2011. New York: IEEE, 1474 – 1479.
35. Hosseinmardi, H., Mattson, S.A., Rafiq, R., Han, R., Lv, Q. and Mishra, S. (2015). Poster: Detection of Cyberbullying in a Mobile Social Network: Systems Issues. IN: Annual International Conference on Mobile Systems, Applications, and Services. 13th. Florence. May 18th – 22nd, 2015. ACM, 481-481.
36. Huang, Q., Singh, V.K. and Atrey, P.K., (2014). Cyberbullying detection using social and textual analysis. IN: International Workshop on Socially-Aware Multimedia. 3rd. Orlando. November 7, 2014. ACM, 3-6.
37. Kleinberg, J.M. (1999). Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM (JACM)*, 46(5), p.604-632.
38. Kontostathis, A. and Pottenger, W.M. (2006). A Framework for Understanding Latent Semantic Indexing (LSI) Performance. *Information Processing & Management*, 42(1), p.56-73.
39. Kontostathis, A., Reynolds, K., Garron, A. and Edwards, L. (2013). Detecting Cyberbullying: Query Terms and Techniques. IN: Annual ACM Web Science Conference. 5th. Indiana. June 23 – 26, 2013. New York: ACM, 195-204.
40. Kovacevic, A. and Nikolic, D., 2014. Automatic Detection of Cyberbullying to Make Internet a Safer Environment. Handbook of Research on Digital Crime, Cyberspace Security, and Information Assurance, p.
41. Kowalski, R.M. and Limber, P. (2007). Electronic Bullying Among Middle School Students. *Journal of Adolescent Health*, 41, p.S22–S30.
42. Kwak, H., Blackburn, J. and Han, S. (2015). Exploring cyberbullying and other toxic behavior in team competition online games. IN: Annual ACM Conference on Human Factors in Computing Systems. 33rd.Seoul. April 18 – 23, 2015. ACM, 3739-3748
43. Li, M. and Tagami, A. (2014). A Study of Contact Network Generation for Cyber-bullying Detection. IN: International Conference on Advanced Information Networking and Applications Workshops (WAINA). 28th. Victoria. May 13-16, 2014. New York: IEEE, 431-436.
44. Li, Q. (2007). New Bottle but Old Wine: A Research of Cyberbullying in Schools. *Computers in human behavior*, 23(4), p.1777-1791.
45. Livingstone, S., Haddon, L., Vincent, J., Mascheroni, G. and Ólafsson, K. (2014). Net Children Go Mobile: The UK Report [online]. Available from http://www.lse.ac.uk/media@lse/research/EUKidsOnline/EU%20Kids%20III/Reports/NCG-MUKReportfinal.pdf [Accessed 21st June 2015].
46. Macbeth, J., Adeyema, H., Lieberman, H. and Fry, C. (2013) Script-based story matching for cyberbullying prevention. IN: CHI'13 Extended Abstracts on Human Factors in Computing Systems. Paris. April 27 - May 02, 2013. ACM, 901-906.
47. Mahmud, A., Ahmed, K.Z. and Khan, M. (2008). Detecting flames and insults in text. Available from http://123.49.46.157/bitstream/handle/10361/714/Detecting%20flames%20and%20insults%20in%20text,%202008.pdf?sequence=1 [Accessed 21st June 2015].
48. Mancilla-Caceres, J., Espelage, D. and Amir, E. (2015). A Computer Game-Based Method for Studying Bullying and Cyberbullying. *Journal of School Violence*, 14(1), 66-86.
49. Mancilla-Caceres, J., Pu, W., Amir, E. and Espelage, D. (2012). A Computer-In-The-Loop Approach For Detecting Bullies In The Classroom. *Social Computing, Behavioral-Cultural Modeling and Prediction*, 7227, p.139 - 146.
50. Mangaonkar, A., Hayrapetian, A. and Raje, R. (2015). Collaborative detection of cyberbullying behavior in Twitter data. IN: IEEE International Conference on Electro/Information Technology (EIT). Illinois. 21 May 21 - 23 May, 2015. IEEE, 611-616.
51. Margono, H., Yi, X. and Raikundalia, G.K. (2014). Mining Indonesian cyber bullying patterns in social networks. IN: Australasian Computer Science Conference. 37th. Auckland, January 20-23, 2014. Australian Computer Society, Inc, 115-124.
52. Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). Efficient estimation of word representations in vector space. Available

from https://arxiv.org/pdf/1301.3781.pdf [Accessed 11th October 2016].

53. Mishna, F., Cook, C., Gadalla, T., Daciuk, J. and Solomon, S. (2010). Cyber Bullying Behaviors among Middle and High School Students. *American Journal of Orthopsychiatry*, 80(3), p.362-374.
54. Mishna, F., Khoury-Kassabri, M., Gadalla, T., and Daciuk, J. (2012). Risk Factors for Involvement in Cyber Bullying: Victims, Bullies and Bully–Victims. *Children and Youth Services Review*, 34(1), p.63-70.
55. Munezero, M., Montero, C.S., Kakkonen, T., Sutinen, E., Mozgovoy, M. and Klyuev, V. (2014). Automatic Detection of Antisocial Behaviour in Texts. *Informatica. Special Issue: Advances in Semantic Information Retrieval*, 38(1), p.3 – 10.
56. Munezero, M., Mozgovoy, M., Kakkonen, T., Klyuev, V. and Sutinen, E. (2013). Antisocial behavior corpus for harmful language detection. IN: Federated Conference on Computer Science and Information Systems (FedCSIS). Krakow. September 8-11, 2013. IEEE, 261-265
57. Nadali, S., Murad, M. A.A., Sharef, N.M., Mustapha, A. and Shojaee, S. (2013). A Review of Cyberbullying Detection: An Overview. IN: International Conference on Intelligent Systems Design and Applications (ISDA). 13th. Malaysia. December 8-10, 2013. New York: IEEE, 325-330.
58. Nahar, V., Al-Maskari, S., Li, X. and Pang, C. (2014). Semi-supervised Learning for Cyberbullying Detection in Social Networks. *Databases Theory and Applications*, 8506, p.160-171.
59. Nahar, V., Li, X. and Pang, C. (2013). An Effective Approach for Cyberbullying Detection. *Communications in Information Science and Management Engineering*, 3(5), p.238.
60. Nahar, V., Unankard, S., Li, X. and Pang, C. (2012). Sentiment Analysis for Effective Detection of Cyber Bullying. Web Technologies and Applications, p.767-774.
61. NaliniPriya. G and Asswini. M. (2015). A dynamic cognitive system for automatic detection and prevention of cyber-bullying attacks. *ARPN Journal of Engineering and Applied Science*, 10(10), pp.4618-4626.
62. Nandhini, B.S. and Sheeba, J.I. (2015a). Online social network bullying detection using intelligence techniques. *Procedia Computer Science*, 45, pp.485-492.
63. Nandhini, B. and Sheeba, J.I. (2015b). Cyberbullying detection and classification using information retrieval algorithm. IN: International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015). Unnao. March 6-7, 2015. ACM, 20.
64. Navarro, J.N. and Jasinski, J. L. (2013). Why Girls? Using Routine Activities Theory to Predict Cyberbullying Experiences between Girls and Boys. *Women & Criminal Justice*, 23(4), p.286-303.
65. Nitta, T., Masui, F., Ptaszynski, M., Kimura, Y., Rzepka, R. and Araki, K. (2013). Detecting Cyberbullying Entries on Informal School Websites Based on Category Relevance Maximization. IN: International Joint Conference on Natural Language Processing (IJCNLP 2013). 6th. Nagoya. 14 – 18, October.
66. Nowak, S. and Rüger, S. (2010). How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation. IN: International Conference on Multimedia Information Retrieval (MIR 2010). 11th. Philadelphia. 29 – 31 March. ACM, 557-566.
67. Oliveira, N., Cortez, P. and Areal, N. (2013). On The Predictability Of Stock Market Behavior Using Stocktwits Sentiment And Posting Volume. *Progress in Artificial Intelligence*, p.355-365.
68. Olweus, D. (1993). Bullying At School: What We Know and What We Can Do. Massachusetts: Wiley-Blackwell.
69. Olweus, D. (2012). Cyberbullying: An Overrated Phenomenon? *European Journal of Developmental Psychology*, 9(5), 520-538.
70. Parime, S. and Suri, V. (2014). Cyberbullying detection and prevention: Data mining and psychological perspective. IN: International Conference on Circuit, Power and Computing Technologies (ICCPCT). Tamil Nadu. March 20 – 21, 2014. IEEE, 1541-1547.
71. Patchin, J.W. and Hinduja, S. (2012). Preventing and Responding To Cyberbullying: Expert Perspectives. Thousand Oaks: Routledge.
72. Pérez, P.J.C., Valdez, C.J.L., Ortiz, M.D.G.C., Barrera, J.P.S. and Pérez, P.F. MISAAC: Instant Messaging Tool for Ciberbullying Detection [online]. Available from http://worldcomp-proceedings.com/proc/p2012/ICA7994.pdf [Accessed 21st June 2015].
73. Potha, N. and Maragoudakis, M. (2014). Cyberbullying detection using time series modeling. IN: IEEE International Conference on Data Mining Workshop (ICDMW). Shenzhen. December 14-17, 2014. IEEE, 373-382.
74. Ptaszynski, M., Dybala, P., Matsuba, T., Masui, F., Rzepka, R. and Araki, K., (2010a). Machine learning and affect analysis against cyber-bullying. AISB Annual Convention. 36th. Leicester. March 29 – April 1, 2010. AISB, 7-16.
75. Ptaszynski, M., Dybala, P., Matsuba, T., Masui, F., Rzepka, R., Araki, K. and Momouchi, Y., (2010b). In the service of online order: Tackling cyber-bullying with machine learning and affect analysis. *International Journal of Computational Linguistics Research*, *1*(3), pp.135-154.
76. Rafiq, R.I., Hosseinmardi, H., Han, R., Lv, Q., Mishra, S. and Mattson, S.A. (2015). Careful what you share in six seconds: detecting cyberbullying instances in Vine. IN: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Paris. August 25-28, 2015. ACM, 617-622
77. Raykar, V. C. and Yu, S. (2012). Eliminating spammers and ranking annotators for crowdsourced labeling tasks. Journal of Machine Learning Research, 13, p.491-518.
78. Sabella, R.A., Patchin, J.W. and Hinduja, S. (2013). Cyberbullying Myths and Realities. *Computers in Human Behavior*, 29(6), p.2703-2711.
79. Saif, H., He, Y. and Alani, H. (2012). Semantic Sentiment Analysis of Twitter. IN: The Semantic Web–ISWC 2012. 11th. Boston. November 11 – 15, 2012. Berlin: Springer, 508-524.
80. Sanchez, H. and Kumar, S. (2011). Twitter Bullying Detection. *NSDI*, 12, p.15-22.
81. Serra, S.M. and Venter, H.S. (2011). Mobile Cyber-Bullying: A Proposal for a Pre-Emptive Approach to Risk Mitigation by Employing Digital Forensic Readiness. *Information Security South Africa (ISSA)*, p.1-5.
82. Ševčíková, A., and Šmahel, D. (2009). Online Harassment and Cyberbullying in the Czech Republic. *Journal of Psychology*, 217(4), 227-229.
83. Sheeba, J.I. and K. Vivekanandan (2013). Low Frequency Keyword Extraction with Sentiment Classification and Cyberbully Detection Using Fuzzy Logic Technique. IN: IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). Enathi. December 26 – 28, 2013. New York: IEEE, 1-5.
84. Sigel, E.J. and Harpin, S.B. (2013). Primary care practitioners' detection of youth violence involvement. *Clinical paediatrics*, 52(5), pp.411-417.
85. Slonje, R. and Smith, P.K. (2007). Cyberbullying: Another Main Type of Bullying? *Scandinavian Journal of Psychology*, 49, p.147–154.
86. Smith, P.K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S. and Tippett, N. (2008). Cyberbullying: Its Nature and Impact in Secondary School Pupils. *Journal of Child Psychology and Psychiatry*, 49(4), p.376-385.
87. Snell, P.A. and Englander, E.K. (2005). Cyberbullying Victimization and Behaviors among Girls: Applying Research Findings in the Field. *Journal of social sciences*, 6(4), p.510 - 514.
88. Sood, S.O., Antin, J. and Churchill, E.F. (2012a). Using Crowdsourcing to Improve Profanity Detection. IN: AAAI Spring Symposium: Wisdom of the Crowd. Stanford, March 26 – 28, 2012. Palo Alto: The AAAI Press, 69 – 74.
89. Sood, S.O., Churchill, E.F. and Antin, J. (2012b). Automatic Identification of Personal Insults on Social News Sites. *Journal of the American Society for Information Science and Technology*, 63(2), p.270-285.
90. Squicciarini, A., Rajtmajer, S., Liu, Y. and Griffin, C. (2015). Identification and characterization of cyberbullying dynamics in an online social network. IN: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Paris. August 25-28, 2015. ACM, 280-285.

91. Snow, R., O'Connor, B., Jurafsky, D. and Ng, A. Y. (2008). Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks. IN: Conference on Empirical Methods in Natural Language Processing. Paris. 25 – 27, October, 2008 Association for Computational Linguistics, 254-263.
92. Strapparava, C. and Valitutti, A. (2004). WordNet Affect: An Affective Extension of WordNet. *LREC*, 4, pp. 1083-1086.
93. Sykes, G.M. and Matza, D. (1957). Techniques of neutralization: A theory of delinquency. *American sociological review*, 22(6), pp.664-670.
94. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D. and Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12), pp.2544-2558.
95. Tokunaga, R. S. (2010). Following You Home From School: A Critical Review and Synthesis of Research on Cyberbullying Victimization. *Computers in Human Behavior*, 26(3), p.277-287.
96. Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., Daelemans, W. and Hoste, V. (2015). Detection and fine-grained classification of cyberbullying events. IN: International Conference Recent Advances in Natural Language Processing (RANLP). Hissar. September 5-11, 2015. RANLP, 672-680.
97. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y. and Manzagol, P. A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, *11*(Dec), 3371-3408.
98. Wang J, Iannotti R. J. and Nansel T.R. (2009) School Bullying Among Adolescents in the United States: Physical, Verbal, Relational, and Cyber. *Journal of Adolescent Health*, 45(4), p.368-375.
99. Welinder, P. and Perona, P. (2010). Online crowdsourcing: Rating annotators and obtaining cost-effective labels. IN: Computer Vision and Pattern Recognition Workshops. San Francisco. June 13 – 18, 2010 IEEE, 25-32. IEEE.
100. William K. and Guerra M. (2007). Prevalence and Predictors of Internet Bullying. *Journal of Adolescent Health*, 41(6), p.S14–S21.
101. Xu, J.M., Jun, K.S., Zhu, X. and Bellmore, A. (2012a). Learning from Bullying Traces in Social Media. IN: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Montreal. June 3 – 8, 2012. Stroudsburg: ACL, 656-666.
102. Xu, J.M., Zhu, X. and Bellmore, A. (2012b). Fast Learning for Sentiment Analysis on Bullying. IN: International Workshop on Issues of Sentiment Discovery and Opinion Mining. 1st. Beijing. August 12 – 16, 2012. New York: ACM.
103. Ybarra, M.L., Mitchell, K.J., Wolak, J. and Finkelhor, D. (2006). Examining Characteristics And Associated Distress Related To Internet Harassment: Findings From The Second Youth Internet Safety Survey. *Paediatrics*, 118, p.e1169–e1177.
104. Yin, D., Xue, Z., Hong, L., Davison, B.D., Kontostathis, A. and Edwards, L. (2009). Detection of Harassment on Web 2.0. IN: Content Analysis in the WEB. Madrid. April 21, 2009.
105. Zhang, S., Yu, L., Wakefield, R.L. and Leidner, D.E., 2016. Friend or Foe: Cyberbullying in Social Network Sites. *ACM SIGMIS Database*, 47(1), pp.51-71.
106. Zhao, R. and Mao, K. (2016) Cyberbullying Detection based on Semantic-Enhanced Marginalized Denoising Auto-Encoder. *IEEE Transactions on Affective Computing*, PP(99), pp.1-12
107. Zhao, R., Zhou, A. and Mao, K. (2016). Automatic detection of cyberbullying on social networks based on bullying features. IN: International Conference on Distributed Computing and Networking. 17th. Singapore. January 04-07, 2016. ACM, 43.

Semiu Salawu is a PhD student at Aston University, UK. He obtained his BTech in Electronics and Electrical Engineering from Ladoke Akintola University of Technology, Nigeria in 2000 and gained MSc and MBA degrees from the University of Wolverhampton, UK in 2006 and 2009 respectively. Semiu has several years' commercial experience as a software engineer and his current research interests lie in machine learning and its application in both commercial and social situations. Semiu is a member of the British Computing Society.

Yulan He is a Reader and Director of the Systems Analytics Research Institute at Aston University, UK. She obtained her BASc (First Class Honours) and MEng degrees in Computer Engineering in 1997 and 2001, respectively, both from Nanyang Technological University, Singapore, and her PhD degree in Spoken Language Understanding in 2004 from the University of Cambridge, UK. Prior joining Aston, she was a Senior Lecturer at the Open University, Lecturer at the University of Exeter and Lecturer at the University of Reading. Her current research interests lie in the integration of machine learning and natural language processing for text mining and social media analysis. Yulan has published over 150 papers in high impact journals and at top conferences such as IEEE Transactions on Knowledge and Data Engineering, IEEE Intelligent Systems, KDD,, ACL, EMNLP, AAAI, etc. She served as an Area Chair in NAACL 2016, EMNLP 2015, CCL 2015 and NLPCC 2015 and co-organized ECIR 2010 and IAPR 2007.

Joanna (Jo) Lumsden obtained a BSc in Software Engineering (Hons) in 1996 and PhD in Human Computer Interaction (HCI) in 2001, both from Glasgow University, UK. She began her research career as a research assistant at Glasgow University, before moving to the National Research Council of Canada (NRC) where she worked as a Research Officer from 2002-2009. She is currently a Reader at Aston University, UK, where is also the Director of the Aston Interactive Media (AIM) Lab. Jo is the Editor in Chief of the International Journal of Mobile Human Computer Interaction (IJMHCI) and serves as a programme committee member/reviewer for numerous leading HCI conferences, including Mobile HCI and CHI, and as an expert reviewer for funding bodies in the UK and Canada. Recipient of several best paper awards, Jo was awarded the Most Influential Paper Award at the 2013 CASCON conference for the influence of her 2003 article in the conference a decade earlier. Jo has published extensively in the field of HCI and has edited several multi-authored books in the field.