information

Article

An Evaluation of Multilingual Offensive Language Identification Methods for the Languages of India

Tharindu Ranasinghe [1] and Marcos Zampieri [2,*]

[1] Research Group in Computational Linguistics, University of Wolverhampton, Wolverhampton WV1 1LY, UK; tharindu.ranasinghe@wlv.ac.uk
[2] Language Technology Group, Rochester Institute of Technology, Rochester, NY 14623, USA
* Correspondence: marcos.zampieri@rit.edu

Abstract: The pervasiveness of offensive content in social media has become an important reason for concern for online platforms. With the aim of improving online safety, a large number of studies applying computational models to identify such content have been published in the last few years, with promising results. The majority of these studies, however, deal with high-resource languages such as English due to the availability of datasets in these languages. Recent work has addressed offensive language identification from a low-resource perspective, exploring data augmentation strategies and trying to take advantage of existing multilingual pretrained models to cope with data scarcity in low-resource scenarios. In this work, we revisit the problem of low-resource offensive language identification by evaluating the performance of multilingual transformers in offensive language identification for languages spoken in India. We investigate languages from different families such as Indo-Aryan (e.g., Bengali, Hindi, and Urdu) and Dravidian (e.g., Tamil, Malayalam, and Kannada), creating important new technology for these languages. The results show that multilingual offensive language identification models perform better than monolingual models and that cross-lingual transformers show strong zero-shot and few-shot performance across languages.

Keywords: offensive language identification; deep learning; multilingual learning

Citation: Ranasinghe, T.; Zampieri, M. An Evaluation of Multilingual Offensive Language Identification Methods for the Languages of India. *Information* **2021**, *12*, 306. https://doi.org/10.3390/info12080306

Academic Editor: Kostas Stefanidis

Received: 25 June 2021
Accepted: 23 July 2021
Published: 29 July 2021

1. Introduction

Computational models trained to identify various types of offensive content online (e.g., hate speech, cyberbullying) have been widely studied in recent years [1]. A number of competitions such as HatEval and OffensEval have been organized, attracting a large number of participants [2,3], which indicates the interest of the AI and NLP communities in this topic. The clear majority of studies in offensive language identification, however, deal with a very small number of high-resource languages, most notably English, due to the availability of large datasets in these languages [4,5]. Taking advantage of recent advances in deep learning representation such as context word embeddings and multilingual transformers in the past several years, a few studies have been published on multilingual models applied to offensive language identification [6–8]. This has opened new avenues for offensive language identification in low-resource languages.

In this paper, we investigate the use of multilingual models to offensive language identification for six languages spoken in India. India is a multilingual country where hundreds of languages are spoken, making it a perfect scenario for multilingual offensive language identification. Furthermore, English is widely spoken in India, and the use of code-mix between English and a local language (e.g., Hindi or Tamil) is pervasive in social media, resulting in a challenging scenario for NLP systems. To the best of our knowledge, this is the first large-scale multilingual study of offensive language identification for the languages of India. We address the question of data scarcity and language similarity/typology, two underexplored issues in offensive language identification. We explore multiple settings with training languages—languages for which we included data when training the

models—and target languages—languages in which we make test set predictions. We explore three main scenarios: (1) zero-shot learning, when a target language does not have any examples; (2) few-shot learning, when a target language has limited training examples, that is, fewer instances than the full training dataset for that language; (3) cross-lingual learning, when the full size target language training set is used regardless of the training set size.

The main contributions of this paper are the following:

1. We applied cross-lingual contextual word embeddings to offensive language identification in six different spoken in India from two language families, Indo-Aryan and Dravidian.
2. We analyzed the feasibility of training a single multilingual model that is able to generalize to multiple languages from different language families.
3. We evaluated the influence of language similarity and typology in cross-lingual offensive language identification by training models using only similar languages.
4. We explored the possibility of using zero-shot and few-shot learning methods for offensive language identification in low-resource languages to address data scarcity with a particular emphasis on language combination.

While transfer learning and multilingual models have become increasingly more popular in NLP in recent years, their use in offensive language identification is still relatively underexplored [8,9]. Furthermore, the few studies recently published in multilingual offensive language identification have used English as a base language to project predictions to various target languages such as German, Hindi, and Spanish. The use of other base languages closely related to the target languages such as Spanish–Portuguese or Hindi–Urdu, however, has not been explored in offensive language identification. To the best of our knowledge, this paper is the first comprehensive study of multilingual offensive language identification models for languages of India, taking into account language similarity by using different base languages and code switching, a common phenomenon in India and a known challenge in NLP.

Motivation

There is a strong need for developing technology to counter harmful online content in India. With the increasing popularity of social media platforms (e.g., Facebook, Twitter) [10] and instant messaging services (e.g., WhatsApp) in India, researchers have been studying the role of phenomena such as online misinformation [11] and hate speech [12] in Indian society and investigating ways to cope with their widespread prevalence.

A major challenge faced by researchers in this area is the lack of resources for most languages spoken in India [13]. India is a linguistically diverse nation and one of the most multilingual countries in the world. While the Indian Constitution recognizes Hindi as the official language of the central government, there are more than 20 official regional languages in India and over a thousand minority languages.

This multilingual scenario creates the need for developing more technology for local languages. This includes offensive language identification systems, which are often modeled as a supervised classification problem relying on large amounts of annotated data [14]. In this paper, we investigate strategies such as zero-shot learning and multilingual learning to circumvent data scarcity in six languages from the two most widely spoken language families in India, namely, Bengali, Hindi, and Urdu, i.e., three Indo-Aryan languages, and Kannada, Malaylam, and Tamil—the three Dravidian languages. As previously stated, to the best of our knowledge, this is the first comprehensive multilingual study of offensive language identification for the languages of India.

Finally, some of the datasets included in this study contain English code-mixed data, which is an important challenge for NLP systems [15]. Given the widespread use of English in a code-mixed setting in India, we believe that our study replicates a real-world scenario common among India speakers, helping to address code-mixed-related challenges in NLP and, more superficially, in the offensive language identification.

2. Related Work

The bulk of work in offensive language identification addressed different types of offensive content such as online abuse [16,17], cyberaggression [1,18], cyberbullying [19], hate speech [20–22], etc. In terms of computational methods, recent work has employed deep neural models such as convolutional neural networks (CNNs) and long, short-term memory (LSTM). With the introduction of transformer-based models, most notably BERT [23], neural transformer models [24] have been widely applied in offensive language identification, topping the leaderboards of competitions such as HatEval [3], HASOC [25], OffensEval [2], and TRAC [18].

The clear majority of these studies have focused on the English language, creating new datasets and resources for this language. Some of the English datasets such as OLID [4] and SOLID [5], used in the popular OffensEval competition at SemEval, have been widely used by the community. A few studies have been published on other languages as well, such as Arabic [26], Greek [27], and Turkish [28], creating important new resources for languages other than English.

To take advantage of available datasets in English, recent studies have explored data augmentation techniques [6], multilingual word embeddings [7], and most recently, cross-lingual contextual word embeddings [8] for low-resource languages, that is, languages with very limited training data available. State-of-the-art cross-lingual contextual embeddings such as XLM-R [29] have been recently applied to offensive language identification, achieving state-of-the-art results for Bengali, Hindi, and Spanish and thus serving as inspiration for this study [8].

Offensive Language Identification in Languages from India

A few recent competitions have provided datasets in multiple languages from India, creating important resources and benchmarks for these languages. These include the aforementioned HASOC shared task, organized from 2019 to 2021, the TRAC shared task, organized in 2018 and 2020, and the shared task on offensive language identification in Dravidian languages at the Dravidian LangTech workshop 2021.

Two iterations of the TRAC shared task on aggression identification have been organized jointly with the TRAC workshop. TRAC 2018 [18] at COLING provided participants with training and test sets containing Facebook comments and a test set containing tweets in Hindi and English. The task was to discriminate between posts labeled as *Aggressive*, *Covertly Aggressive*, and *Non-aggressive*. In terms of performance, systems using traditional machine learning classifiers such as SVMs performed at par with neural network-based systems [18]. TRAC 2020 [1] at LREC provided participants with Bengali, English, and Hindi datasets containing YouTube comments. Two subtasks were included—subtask A contained the same three classes as TRAC 2018, whereas subtask B contained two classes, one of which aimed to identify gendered aggression in posts targeted at women. In terms of performance, a system based on pretrained transformer models such as BERT performed best [30].

The HASOC shared task, which stands for "hate speech and offensive content identification", in Indo-European Languages is arguably the most well-known series of competitions including languages from India [25,31]. It has been organized in 2019 and 2020 at the Forum for Information Retrieval (FIRE). HASOC 2019 provided participants with datasets in English, German, and Hindi, while HASOC 2020 featured the aforementioned three languages plus Tamil and Malayalam. In terms of performance, systems based on neural network architectures have been shown to achieve competitive performance [24]. HASOC 2021 is currently ongoing with the addition of Marathi.

The shared task at Dravidian LangTech [32] focused on identifying offensive language content of the code-mixed dataset of comments and posts in three Dravidian Languages, namely, Tamil–English, Malayalam–English, and Kannada–English collected from social media. These three Dravidian languages are closely related, presenting us with a good opportunity to use multilingual models for offensive language identification on these data,

but at the same time, the similarity between these languages often pose challenges for NLP pipelines as explored in the recent Dravidian language identification (DLI) shared task at VarDial [33]. In Dravidian LangTech, most of the top-performing systems [34–36] used neural network architectures based on pretrained transformer models such as multilingual BERT [23], XLM-R [29], and Indic-BERT [37]. However, none of them considered performing transfer learning from different languages to improve the performance. The Tamil–English and the Kannada–English datasets that were used in this research are taken from this shared task.

Only a limited number of studies have been conducted on the impact of transfer learning for offensive language identification in languages from India. Drawing inspiration from Ranasinghe and Zampieri [8], Sai and Sharma [9] improve offensive language identification for code-mixed Kannada, Malayalam, and Tamil by performing transfer learning from the English OLID dataset [4]. On a different research, Ranasinghe et al. [38] improve offensive language identification for code-mixed Malayalam using transfer learning from English. However, both of these papers only considered transfer learning from English, which leaves a considerable space to explore transfer learning within different languages in India. Furthermore, to the best of our knowledge, there were no transfer learning studies published on transferring between languages from India. Our work fills this important gap, opening new avenues for future research for multiple languages from India.

3. Data

As data for this research, we considered six different native languages that are very popular in India—Bengali, Hindi, Kannada, Malayalam, Tamil, and Urdu. Other than these native languages, we also considered English, which is widely used in India. We used nine recently released offensive language identification datasets in these languages collected from Twitter and YouTube. Detailed information on these languages are provided in Table 1.

Table 1. Instances (Inst.), source (S), and labels in all datasets. T stands for Twitter and Y for Youtube. In the Kannada and Tamil datasets, the label "Offensive Target-Insult" is further subdivided into the OLID level C categories.

Language	I	S	Labels	Family
English [4]	14,100	T	offensive, non-offensive	
Bengali [39]	4000	Y	overtly aggressive, covertly aggressive, non-aggressive	Indo-Aryan
Hindi [31]	8000	T	hate offensive, non-hate-offensive	Indo-Aryan
Hindi–English [40]	4114	T	hate Speech, normal Speech	Indo-Aryan
Kannada–English [41]	7671	Y	Non-offensive, Offensive-untargeted, Offensive Targeted-Insult	Dravidian
Malayalam–English [25]	4000	T	offensive, not offensive	Dravidian
Tamil–English [42]	8000	Y	not-offensive, offensive-untargeted, offensive Targeted-Insult	Dravidian
Urdu [43]	2171	T	offensive, non-offensive	Indo-Aryan
Urdu–English [43]	10,000	T	offensive, non-offensive	Indo-Aryan

As can be seen in Table 1, the largest dataset included in this study is in English, with over 14,000 instances. This once again confirms that also in our study, English is the language with the most number of resources, while the datasets available for the languages of India are smaller or, in the case of Urdu, much smaller. Furthermore, it should be noted that the datasets listed as Hindi–English, Kannada–English, Malayalam–English, Tamil–English, and Urdu–English contain code-mixed instances, a known challenge for NLP applications and a particularly relevant one considering the linguistic situation of India.

In terms of their annotation, the majority of the offensive language identification datasets we considered have been annotated using only two labels—offensive and non-offensive. In order to perform transfer learning and zero-shot learning across languages, it is paramount to have the same number of labels in all the datasets. Therefore, we mapped the classes of the other datasets that have more than two labels into the offensive vs. non-offensive distinction presented in OLID [4], one of the most widely used English offensive language identification datasets and the dataset we used in this paper. For Bengali [39], we concatenated overtly aggressive and covertly aggressive labels to make a single aggressive label, and the alternated dataset will have only two labels aggressive and non-aggressive. For Kannada–English [41] and Tamil–English [42] datasets, we concatenated Offensive-untargeted and offensive targeted-insult labels to create a single offensive label so that the alternated dataset would have only two labels, non-offensive and offensive. Furthermore, in the Kannada–English and Tamil–English datasets, a label not-Kannada and not-Tamil are included for comments that are not from these two languages. We discarded those comments in our experiments.

4. Architecture

Since this research is motivated by multilingualism, we have considered different multilingual pretrained transformer models for our text classification architecture. Even though there were several multilingual models such as BERT-m [23], there are many speculations about its ability to represent all the languages [44,45]. Although the BERT-m model showed some cross-lingual characteristics, it should be noted that it has not been trained on cross-lingual data [46]. On the other hand, XLM-R [29] has been trained on a huge, multilingual dataset at an enormous scale: unlabeled text in 104 languages, totaling 2.5TB, is extracted from the CommonCrawl datasets. It is trained using only RoBERTa's [47] masked language modeling (MLM) objective [29]. Surprisingly, this strategy provided better results in cross-lingual tasks. XLM-R outperforms mBERT on a variety of cross-lingual benchmarks such as cross-lingual natural language inference and cross-lingual question answering [29]. As we mentioned before, the cross-lingual nature of XLM-R has proven to be advantageous in previous multilingual offensive language research [8]. Therefore, our architecture relied on the XLM-R transformer model [29] to derive the representations of the input sentences.

Similar to other transformer architectures, XLM-R transformer architecture can also be used for text classification tasks [29]. The XLM-R-large model contains approximately 125 million parameters with 12-layers, 768 hidden states, 3072 feed-forward hidden states, and 8 heads [29]. It takes an input of a sequence of no more than 512 tokens and outputs the representation of the sequence. The first token of the sequence is always [CLS], which contains the special classification embedding [48].

For text classification tasks, XLM-R takes the final hidden state h of the first token [CLS] as the representation of the whole sequence. A simple softmax classifier is added to the top of XLM-R to predict the probability of label c: as shown in Equation (1), where W is the task-specific parameter matrix [49,50].

$$p(c|\mathbf{h}) = softmax(W\mathbf{h}) \tag{1}$$

In the classification task, all the parameters from XLM-R as well as W fine-tuned jointly by maximizing the log probability of the correct label. The architecture diagram of the model is shown in Figure 1.

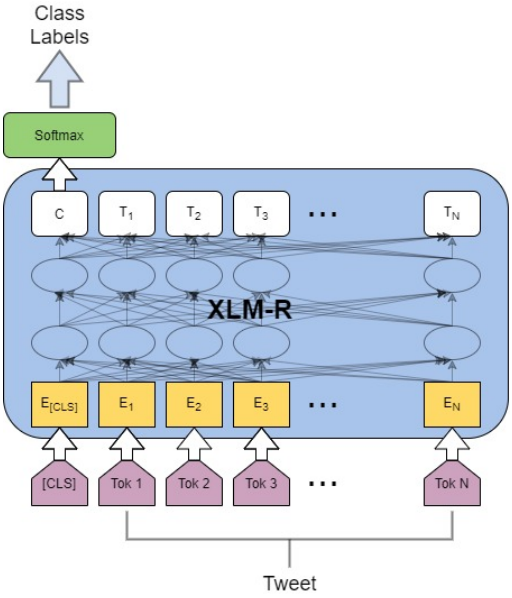


Figure 1. Text classification architecture [8].

5. Experimental Setup

5.1. Running Configurations

We used an Nvidia Tesla K80 GPU to train the models. We divided the dataset into a training set and a validation set using a 0.8:0.2 split on the dataset. We mainly fine-tuned the learning rate and a number of epochs of the classification model manually to obtain the best results for the validation set in each language. We obtained 1×10^{-5} as the best value for learning rate and 3 as the best value for a number of epochs for all the languages. The other configurations of the transformer model were set to a constant value over all the languages in order to ensure consistency between the languages. We used a batch size of eight, an Adam optimizer, and a linear learning rate warm-up of over 10% of the training data. The models were trained using only training data. We performed early stopping if the evaluation loss did not improve over 10 evaluation rounds. A summary of hyperparameters and their values used to obtain the reported results are mentioned in Table 2. The optimized hyperparameters are marked with ‡, and their optimal values are reported. The rest of the hyperparameter values are kept as constants.

Table 2. Hyperparameter specifications.

Parameter	Value
learning rate ‡	1×10^{-5}
number of epochs ‡	3
adam epsilon	1×10^{-8}
warmup ration	0.1
warmup steps	0
max grad norm	1.0
max seq. length	120
gradient accumulation steps	1

5.2. Evaluation Method

Given the strong imbalance between the number of instances in the offensive class and non-offensive class, we used the macro-averaged F1 score shown in Equation (2) as the evaluation measure for all the languages, which has been used in recent OffensEval tasks [2,51].

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \quad (2)$$

6. Results

We first evaluated our architecture in a supervised monolingual setting where the model was trained on the training set of a particular language and tested on the test set of the same language. In Table 3, we show the results comparing our architecture with the best systems and baselines in each dataset. For comparability purposes, we only show the results for the datasets we did not alter. Additionally, the values of the diagonal of section I of Table 4 show the results for all the languages.

Table 3. Results ordered by Macro F1 for English, Hindi, Hindi–English, Malayalam–English, Urdu, and Urdu–English datasets.

Language	Model	Macro F1
English	Wiedemann et al. [52]	0.9204
	XLM-R	0.9123
	Zampieri et al. [4]	0.8000
	Majority Baseline	0.4193
Hindi	Bashar and Nayak [53]	0.8149
	XLM-R	0.8061
	Majority Baseline [31]	0.3510
Hindi–English	XLM-R	0.7782
	Bohra et al. [40]	0.7170
	Majority Baseline [40]	0.4542
Malayalam–English	Sai and Sharma [54]	0.9504
	XLM-R [38]	0.9332
	Majority Baseline [25]	0.7652
Urdu	Akhter et al. [43]	0.9591
	XLM-R	0.9503
	Majority Baseline	0.5000
Urdu–English	Akhter et al. [43]	0.9901
	XLM-R	0.9891
	Majority Baseline	0.4842

As can be seen in Table 3 the architecture performs on par with the best systems available in all the languages and even outperforms it in Hindi–English. It should be noted that usually, the best systems have been built using monolingual embeddings Zampieri et al. [4], which normally outperforms multilingual embeddings in that particular language [23]. Therefore, we believe that the results of XLM-R being lower than the best system are expected. Despite this fact, XLM-R is still very compatible across all the languages. In the following sections, we examine its behavior in different settings.

6.1. Multilingual Offensive Language Identification

We combined instances from all the languages and built a single offensive language identification model. Our results, displayed in section III ("All") of Table 4, show that multilingual models perform better than monolingual models for all the languages in offensive language identification. We believe that cross-lingual transformer models benefits from the advantage of having more data to fine-tune its weights better.

We also investigated whether combining languages that are from the same language group can be more beneficial since it is possible that the learning process is better when languages share certain characteristics. Section II of Table 4 shows these results. Results show that language-group specific models perform better than monolingual models and perform slightly better than multilingual models in all the languages we considered. We believe that the learning process of the transformer model becomes easier when the languages are from the same group. Therefore, the offensive language identification models

built on a specific language group perform slightly better than the purely multilingual offensive language identification models.

Table 4. Macro-average F1 between the algorithm predictions and human annotations. Best results for each language by any method are marked in bold. Sections I, II, and III indicate the different evaluation settings. Zero-shot results are colored in grey and it shows the difference between the best result in that section for that language pair and itself.

	Train Language(s)	Bengali	English	Hindi	Hindi–English	Kannada–English	Malayalam–English	Tamil–English	Urdu	Urdu–English
I	Bengali	0.8751	(−0.09)	(−0.07)	(−0.09)	(−0.10)	(−0.10)	(−0.09)	(−0.07)	(−0.07)
	English	(−0.12)	0.9123	(−0.11)	(−0.08)	(−0.07)	(−0.07)	(−0.08)	(−0.11)	(−0.07)
	Hindi	(−0.07)	(−0.10)	0.8061	(−0.03)	(−0.08)	(−0.08)	(−0.09)	(−0.04)	(−0.05)
	Hindi–English	(−0.11)	(−0.05)	(−0.04)	0.7782	(−0.06)	(−0.06)	(−0.07)	(−0.05)	(−0.05)
	Kannada–English	(−0.12)	(−0.06)	(−0.11)	(−0.08)	0.8153	(−0.06)	(−0.06)	(−0.11)	(−0.09)
	Malayalam–English	(−0.13)	(−0.07)	(−0.12)	(−0.08)	(−0.07)	0.9332	(−0.07)	(0.11)	(−0.08)
	Tamil–English	(−0.04)	(−0.08)	(−0.10)	(−0.09)	(−0.16)	(−0.01)	0.8334	(−0.15)	(−0.14)
	Urdu	(−0.09)	(−0.09)	(−0.08)	(−0.09)	(−0.14)	(−0.12)	(−0.11)	0.9503	(−0.04)
	Urdu–English	(−0.09)	(−0.05)	(−0.08)	(−0.07)	(−0.12)	(−0.11)	(−0.13)	(−0.03)	0.9891
II	Language Group	**0.8892**	NA	**0.8334**	**0.7981**	**0.8341**	**0.9487**	**0.8568**	**0.9698**	**0.9911**
III	All-1	(−0.02)	(−0.02)	(−0.02)	(−0.02)	(−0.02)	(−0.02)	(−0.02)	(−0.02)	(−0.03)
	All	0.8853	**0.9120**	0.8225	0.7882	0.8201	0.9412	0.8452	0.9661	0.9900

6.2. Zero-Shot Offensive Language Identification

To test whether an offensive language identification model trained on a particular language can be generalized to other languages, we performed zero-shot offensive language identification. We used the offensive language identification model trained on a particular language and extended it to the test sets of the other languages. Non-diagonal values of section I in Table 4 shows how each offensive language identification model performed on other languages. For better visualization, the non-diagonal values of section I of Table 4 show how much the score changes when the zero-shot offensive language identification model is used instead of the monolingual offensive language identification model. As can be seen, the scores decrease, but this decrease is small and to be expected. The results show some interesting patterns between languages when performing zero-shot offensive language identification, which include the following:

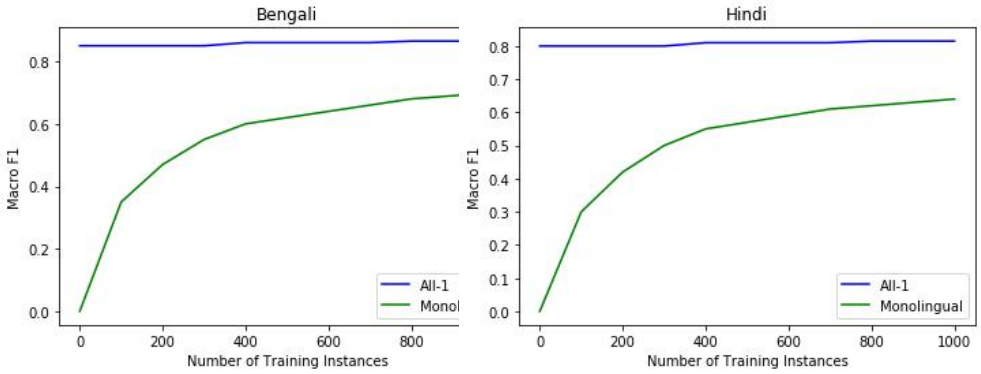
1. Performing zero-shot learning for a code-switched dataset is better when the trained model is based on English or that particular language. For example, zero-shot results on Hindi–English are better when you perform transfer learning from Hindi or English rather than a completely different language such as Bengali or Urdu.
2. A model trained on code-mixed data on a particular language is better for zero-shot learning in not code-mixed data in that particular language. For example, performing zero-shot learning from Hindi–English to Hindi is better than performing zero-shot learning from Bengali to Hindi.
3. Performing zero-shot learning is better inside the language groups. For example, performing zero-shot learning from Hindi to Urdu is better than performing zero-shot learning from Hindi to English–Tamil or English—Kannada since both Hindi and Urdu belong to the same language pair.

We also experimented with zero-shot offensive language identification with multilingual offensive language identification models. We trained the offensive language identification model in all the languages, except one, and performed prediction on the test set of the language left out. In section II ("All-1"), we show its differences from the multilingual offensive language

identification model. This also provides competitive results for the majority of the languages, proving that it is possible to train a single multilingual offensive language identification model and extend it to a multiple languages. This approach provides better results than performing transfer learning from a monolingual model. Therefore, we can assume that a model trained in multiple languages has better knowledge and can perform better than monolingual models in zero-shot learning.

6.3. Few-Shot Offensive Language Identification

In order to examine how the multilingual offensive language identification model performs in a few-shot scenario for an unseen language, we performed few-shot learning. For a particular language, we took its relevant "All-1" model, used its weights to initialize the training, and performed training only on a limited number of training examples. We compared this to training from scratch to the same number of training examples. The results for Bengali and Hindi are shown in Figure 2. As shown in the graph, multilingual models clearly outperform monolingual models in few-shot learning. Since it follows the same trend in all the other languages as well, we did not include them in the graph. From the results, we can state that when a particular language has a few training instances, it is better to transfer weights from a multilingual model and perform training rather than building a monolingual model from scratch.



(**a**) Bengali Results (**b**) Hindi Results

Figure 2. Transfer learning impact in offensive language identification for Bengali and Hindi. Monolingual indicates that the model was trained from scratch, while All-1 indicates that the model was trained on all the languages except that language.

7. Conclusions

In this paper, we explored multilingual offensive language identification with transformers for six languages spoken in India. In our experiments, we observed that multilingual offensive language identification models provide strong results on the language pairs they were trained in. In addition, the multilingual offensive language identification models perform well in the majority of the zero-shot scenarios where the multilingual offensive language identification model is tested on an unseen language. The results confirm the feasibility of training a single multilingual offensive language identification model in as many languages as possible and then applying it to other unseen languages. We believe this outcome opens exciting new avenues in multilingual offensive language identification.

Furthermore, when there is only a limited number of training instances available, our results show that it is better to perform transfer learning from a multilingual model rather than building a monolingual model from scratch. The lessons learned in our experiments are an important contribution in offensive language identification in low-resource languages where training data are scarce and when maintaining several offensive language identification models for different languages is arduous.

As future research, we would like to incorporate more regional languages of India into the multilingual offensive language identification model. As discussed in this paper, there are over 20 official regional languages in India for which very few resources are available, making them an ideal candidate for the cross-lingual models evaluated in this paper. Finally, we extend our research to very low-resource languages that XLM-R pretrained model does not support and examine how the cross-lingual transformer model solves the offensive language identification by gaining knowledge from similar languages that exist in XLM-R.

Author Contributions: Conceptualization, T.R. and M.Z.; methodology, T.R. and M.Z.; software, T.R.; validation, T.R.; formal analysis, T.R. and M.Z.; investigation, T.R. and M.Z.; resources, T.R. and M.Z.; data curation, T.R. and M.Z.; writing—original draft preparation, T.R. and M.Z.; writing—review and editing, T.R. and M.Z.; visualization, T.R. and M.Z.; supervision, M.Z.; project administration, M.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets are made available by their respective data providers. A copy is available on this GitHub repository: https://github.com/TharinduDR/DeepOffense/tree/master/examples (accessed on 28 July 2021).

Acknowledgments: We would like to thank the authors of all datasets used in this paper for making the data available to the community. We further thank the anonymous Information reviewers for their insightful feedback.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kumar, R.; Ojha, A.K.; Malmasi, S.; Zampieri, M. Evaluating Aggression Identification in Social Media. In Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, Marseille, France, 16 May 2020; pp. 1–5.
2. Zampieri, M.; Malmasi, S.; Nakov, P.; Rosenthal, S.; Farra, N.; Kumar, R. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 75–86. [CrossRef]
3. Basile, V.; Bosco, C.; Fersini, E.; Nozza, D.; Patti, V.; Rangel Pardo, F.M.; Rosso, P.; Sanguinetti, M. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 54–63. [CrossRef]
4. Zampieri, M.; Malmasi, S.; Nakov, P.; Rosenthal, S.; Farra, N.; Kumar, R. Predicting the Type and Target of Offensive Posts in Social Media. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 1415–1420. [CrossRef]
5. Rosenthal, S.; Atanasova, P.; Karadzhov, G.; Zampieri, M.; Nakov, P. SOLID: A Large-Scale Semi-Supervised Dataset for Offensive Language Identification. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, 17 July 2021; pp. 915–928. [CrossRef]
6. Ghadery, E.; Moens, M.F. LIIR at SemEval-2020 Task 12: A Cross-Lingual Augmentation Approach for Multilingual Offensive Language Identification. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, Online, 12–13 December 2020; pp. 2073–2079.
7. Pamungkas, E.W.; Patti, V. Cross-domain and Cross-lingual Abusive Language Detection: A Hybrid Approach with Deep Learning and a Multilingual Lexicon. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Florence, Italy, 29–31 July 2019; pp. 363–370. [CrossRef]
8. Ranasinghe, T.; Zampieri, M. Multilingual Offensive Language Identification with Cross-lingual Embeddings. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 5838–5844. [CrossRef]
9. Sai, S.; Sharma, Y. Towards Offensive Language Identification for Dravidian Languages. In Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Kyiv, Ukraine, 19–20 April 2021; pp. 18–27.
10. Hettiarachchi, H.; Adedoyin-Olowe, M.; Bhogal, J.; Gaber, M.M. Embed2Detect: Temporally clustered embedded words for event detection in social media. *Mach. Learn.* **2021**. [CrossRef]

11. Akbar, S.Z.; Panda, A.; Kukreti, D.; Meena, A.; Pal, J. Misinformation as a Window into Prejudice: COVID-19 and the Information Environment in India. *Proc. ACM Hum.-Comput. Interact.* **2021**, *4*. [CrossRef]
12. Sharma, I. Contextualising Hate Speech: A Study of India And Malaysia. *Millenn. J. Int. Stud.* **2019**, *15*, 133–144. [CrossRef]
13. Ranasinghe, T.; Zampieri, M. Multilingual Offensive Language Identification for Low-resource Languages. *arXiv* **2021**, arXiv:2105.05996.
14. Ranasinghe, T.; Zampieri, M. MUDES: Multilingual Detection of Offensive Spans. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations, Online, 8–9 June 2021; pp. 144–152. [CrossRef]
15. Solorio, T.; Blair, E.; Maharjan, S.; Bethard, S.; Diab, M.; Ghoneim, M.; Hawwari, A.; AlGhamdi, F.; Hirschberg, J.; Chang, A.; et al. Overview for the First Shared Task on Language Identification in Code-Switched Data. In Proceedings of the First Workshop on Computational Approaches to Code Switching, Doha, Qatar, 25 October 2014; pp. 62–72. [CrossRef]
16. Mubarak, H.; Darwish, K.; Magdy, W. Abusive Language Detection on Arabic Social Media. In Proceedings of the First Workshop on Abusive Language, Vancouver, BC, Canada, 4 August 2017; pp. 52–56. [CrossRef]
17. Vidgen, B.; Derczynski, L. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS ONE* **2021**, *15*, e0243300. [CrossRef]
18. Kumar, R.; Ojha, A.K.; Malmasi, S.; Zampieri, M. Benchmarking Aggression Identification in Social Media. In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), Santa Fe, NM, USA, 25 August 2018; pp. 1–11.
19. Rosa, H.; Pereira, N.; Ribeiro, R.; Ferreira, P.C.; Carvalho, J.P.; Oliveira, S.; Coheur, L.; Paulino, P.; Simão, A.V.; Trancoso, I. Automatic cyberbullying detection: A systematic review. *Comput. Hum. Behav.* **2019**, *93*, 333–345. [CrossRef]
20. Malmasi, S.; Zampieri, M. Detecting Hate Speech in Social Media. In Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, 2–8 September 2017; pp. 467–472. [CrossRef]
21. Malmasi, S.; Zampieri, M. Challenges in Discriminating Profanity from Hate Speech. *J. Exp. Theor. Artif. Intell.* **2018**, *30*, 1–16. [CrossRef]
22. Röttger, P.; Vidgen, B.; Nguyen, D.; Waseem, Z.; Margetts, H.; Pierrehumbert, J. HateCheck: Functional Tests for Hate Speech Detection Models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online, 17 July 2021; pp. 41–58. [CrossRef]
23. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. [CrossRef]
24. Ranasinghe, T.; Zampieri, M.; Hettiarachchi, H. BRUMS at HASOC 2019: Deep Learning Models for Multilingual Hate Speech and Offensive Language Identification. In Proceedings of the 11th Forum for Information Retrieval, Kolkata, India, 12–15 December 2019.
25. Mandl, T.; Modha, S.; Kumar M, A.; Chakravarthi, B.R. Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German. In Proceedings of the FIRE 2020: Forum for Information Retrieval Evaluation, Hyderabad, India, 16–20 December 2020; pp. 29–32. [CrossRef]
26. Mubarak, H.; Rashed, A.; Darwish, K.; Samih, Y.; Abdelali, A. Arabic Offensive Language on Twitter: Analysis and Experiments. In Proceedings of the Sixth Arabic Natural Language Processing Workshop (Virtual), Kyiv, Ukraine, 19 April 2021; pp. 126–135.
27. Pitenis, Z.; Zampieri, M.; Ranasinghe, T. Offensive Language Identification in Greek. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 5113–5119.
28. Çöltekin, Ç. A Corpus of Turkish Offensive Language on Social Media. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 6174–6184.
29. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 8440–8451. [CrossRef]
30. Risch, J.; Krestel, R. Bagging BERT Models for Robust Aggression Identification. In Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, Marseille, France, 16 May 2020; pp. 55–61.
31. Mandl, T.; Modha, S.; Majumder, P.; Patel, D.; Dave, M.; Mandlia, C.; Patel, A. Overview of the HASOC Track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages. In Proceedings of the 11th Forum for Information Retrieval Evaluation, Kolkata, India, 12–15 December 2019; pp. 14–17. [CrossRef]
32. Chakravarthi, B.R.; Priyadharshini, R.; Jose, N.; Mandl, T.; Kumaresan, P.K.; Ponnusamy, R.; Hariharan, R.L.; McCrae, J.P.; Sherly, E. Findings of the Shared Task on Offensive Language Identification in Tamil, Malayalam, and Kannada. In Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Kyiv, Ukraine, 19 April 2021; pp. 133–145.
33. Chakravarthi, B.R.; Mihaela, G.; Ionescu, R.T.; Jauhiainen, H.; Jauhiainen, T.; Lindén, K.; Ljubešić, N.; Partanen, N.; Priyadharshini, R.; Purschke, C.; et al. Findings of the VarDial Evaluation Campaign 2021. In Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects, Kiyv, Ukraine, 20 April 2021; pp. 1–11.
34. Saha, D.; Paharia, N.; Chakraborty, D.; Saha, P.; Mukherjee, A. Hate-Alert@DravidianLangTech-EACL2021: Ensembling strategies for Transformer-based Offensive language Detection. In Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Kyiv, Ukraine, 19 April 2021; pp. 270–276.

35. Kedia, K.; Nandy, A. indicnlp@kgp at DravidianLangTech-EACL2021: Offensive Language Identification in Dravidian Languages. In Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Kyiv, Ukraine, 19 April 2021; pp. 330–335.
36. Balouchzahi, F.; Aparna, B.K.; Shashirekha, H.L. MUCS@DravidianLangTech-EACL2021:COOLI-Code-Mixing Offensive Language Identification. In Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Kyiv, Ukraine, 19 April 2021; pp. 323–329.
37. Kakwani, D.; Kunchukuttan, A.; Golla, S.; Gokul, N.C.; Bhattacharyya, A.; Khapra, M.M.; Kumar, P. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 16–20 November 2020; pp. 4948–4961. [CrossRef]
38. Ranasinghe, T.; Gupte, S.; Zampieri, M.; Nwogu, I. WLV-RIT at HASOC-Dravidian-CodeMix-FIRE2020: Offensive Language Identification in Code-switched YouTube Comments. In Proceedings of the 12th Forum for Information Retrieval, Hyderabad, India, 16–20 December 2020.
39. Bhattacharya, S.; Singh, S.; Kumar, R.; Bansal, A.; Bhagat, A.; Dawer, Y.; Lahiri, B.; Ojha, A.K. Developing a Multilingual Annotated Corpus of Misogyny and Aggression. In Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, Marseille, France, 16 May 2020; pp. 158–168.
40. Bohra, A.; Vijay, D.; Singh, V.; Akhtar, S.S.; Shrivastava, M. A Dataset of Hindi-English Code-Mixed Social Media Text for Hate Speech Detection. In Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, New Orleans, LA, USA, 6 June 2018; pp. 36–41. [CrossRef]
41. Hande, A.; Priyadharshini, R.; Chakravarthi, B.R. KanCMD: Kannada CodeMixed Dataset for Sentiment Analysis and Offensive Language Detection. In Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media, Online, 13 December 2020; pp. 54–63.
42. Chakravarthi, B.R.; Muralidaran, V.; Priyadharshini, R.; McCrae, J.P. Corpus Creation for Sentiment Analysis in Code-Mixed Tamil-English Text. In Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), Marseille, France, 11 May 2020; pp. 202–210.
43. Akhter, M.P.; Jiangbin, Z.; Naqvi, I.R.; Abdelmajeed, M.; Sadiq, M.T. Automatic Detection of Offensive Language for Urdu and Roman Urdu. *IEEE Access* **2020**, *8*, 91213–91226. [CrossRef]
44. Pires, T.; Schlinger, E.; Garrette, D. How Multilingual is Multilingual BERT? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 4996–5001. [CrossRef]
45. Uyangodage, L.; Ranasinghe, T.; Hettiarachchi, H. Transformers to Fight the COVID-19 Infodemic. In Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, Online, 6 June 2021; pp. 130–135. [CrossRef]
46. K, K.; Wang, Z.; Mayhew, S.; Roth, D. Cross-Lingual Ability of Multilingual BERT: An Empirical Study. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 1 May–26 April 2020.
47. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
48. Sun, C.; Qiu, X.; Xu, Y.; Huang, X. *How to Fine-Tune BERT for Text Classification? Chinese Computational Linguistics*; Sun, M., Huang, X., Ji, H., Liu, Z., Liu, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 194–206.
49. Ranasinghe, T.; Hettiarachchi, H. BRUMS at SemEval-2020 Task 12: Transformer Based Multilingual Offensive Language Identification in Social Media. In Proceedings of the Fourteenth Workshop on Semantic Evaluation (Online), Barcelona, Spain, 12–13 December 2020; pp. 1906–1915.
50. Hettiarachchi, H.; Ranasinghe, T. InfoMiner at WNUT-2020 Task 2: Transformer-based Covid-19 Informative Tweet Extraction. In Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020), Online, 19 November 2020; pp. 359–365. [CrossRef]
51. Zampieri, M.; Nakov, P.; Rosenthal, S.; Atanasova, P.; Karadzhov, G.; Mubarak, H.; Derczynski, L.; Pitenis, Z.; Çöltekin, Ç. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In Proceedings of the Fourteenth Workshop on Semantic Evaluation, Online, 12–13 December 2020; pp. 1425–1447.
52. Wiedemann, G.; Yimam, S.M.; Biemann, C. UHH-LT at SemEval-2020 Task 12: Fine-Tuning of Pre-Trained Transformer Networks for Offensive Language Detection. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, Online, 12–13 December 2020; pp. 1638–1644.
53. Bashar, M.A.; Nayak, R. QutNocturnal@ HASOC'19: CNN for hate speech and offensive content identification in Hindi language. In Proceedings of the 11th Forum for Information Retrieval, Kolkata, India, 12–15 December 2019.
54. Sai, S.; Sharma, Y. Siva@HASOC-Dravidian-CodeMix-FIRE-2020: Multilingual Offensive Speech Detection in Code-mixed and Romanized Text. In Proceedings of the 12th Forum for Information Retrieval, Hyderabad, India, 16–20 December 2020.