A Transformer Based Approach for Abuse Detection in Code Mixed Indic Languages.

VIBHUTI BANSAL and MRINAL TYAGI, Bharati Vidyapeeth's College of Engineering, India
RAJESH SHARMA, Institute of Computer Science, University of Tartu, Estonia
VEDIKA GUPTA, Jindal Global Business School, O.P. Jindal Global University, India
QIN XIN, Faculty of Science and Technology, University of the Faroe Islands, Faroe Islands

The advancement in the number of online social media platforms has entailed active participation from the web users globally. This has also lead to subsequent increase in the cyberbullying cases online. Such incidents diminish an individual's reputation or defame a community, also posing a threat to the privacy of users in cyberspace. Traditionally, manual checks and handling mechanisms have been used to deal with such textual content. However, an automatic computer-based approach would provide far better solutions to this problem. Existing approaches to automate this task majorly involves classical machine learning models which tend to perform poorly on low resource languages. Owing to the varied background and language of web users, the cyberspace witnesses the presence of multilingual text. An integrated approach to accommodate multilingual text could be the appropriate solution. This paper explores various methods to detect abusive content in 13 Indic code-mixed languages. Firstly, baseline classical machine learning models are compared with Transformer based architecture. Secondly, the paper presents the experimental analysis of four state-of-the-art transformer-based models vis à vis XLM-RoBERTa, indic-BERT, MurilBert and mBERT, out of which XLM Roberta with BiGRU outperforms. Thirdly, the experimental setup of the best performing model XLM-RoBERTa is fed with emoji embeddings that leads to further enhancement of overall performance of the employed model. Finally, the model is trained with the combined dataset of 13 Indic languages, to compare its performance with those of individual language models. The performance of combined model surpassed those of the individual models in terms of F1 score and accuracy, supporting the fact that combined model fits the data better possibly due to its code-mixed nature. This model reports a F1 score of 0.88 on test data while rendering a training loss of 0.28, validation loss of 0.31 and an AUC score of 0.94 for both training and validation.

CCS Concepts: • **Information systems** → Presentation of retrieval results; Social networking sites; Information Retrieval; Structure and multilingual text search.

Additional Key Words and Phrases: Abuse detection, Transformer based model, online social media, machine learning

1 INTRODUCTION

The growing digital dominance in human life can be well-attributed by the parallelly growing numbers of social media platforms as well as Web users. According to statista [1], there were 4.66 billion active internet users globally, accounting for 60% of the total world's population as in January 2021. Social media platforms, in the form of social media applications are easily accessible on personal digital assistants (PDAs) and hand-held devices, providing on-the-go access. An increasing number of social networks are therefore accessible through multiple platforms

[1] https://www.statista.com/statistics/617136/digital-population-worldwide/

Authors' addresses: Vibhuti Bansal, bansal.vibhuti25@gmail.com; Mrinal Tyagi, mrinaltyagi24@gmail.com, Bharati Vidyapeeth's College of Engineering, India; Rajesh Sharma, Institute of Computer Science, University of Tartu, Estonia, rajesh.sharma@ut.ee; Vedika Gupta, Jindal Global Business School, O.P. Jindal Global University, Sonipat, Haryana-131001, India, vgupta2@jgu.edu.in; Qin Xin, Faculty of Science and Technology, University of the Faroe Islands, Vestarabryggja 15, FO 100 Torshavn, Faroe Islands, qinx@setur.fo.

2375-4699/2022/11-ART $15.00
https://doi.org/10.1145/3571818

to offer users access to different features according to their needs, time and preferred device. In the past when the World Wide Web started growing its paws, there have been times where people actively discussed about the pros and cons of Internet [5] [25]. As time progressed, Internet made its mark and proved to be of significant advantage and such debatable conversations gradually declined. Of course, the Internet today has grown more advantageous and brought in plethora of possibilities for mankind; nonetheless, it has outreached our capabilities of limiting or scrutinizing every piece of information that is shared or posted online.

Social media is an important tool to reach to masses for visibility of ideas and sharing thoughts but at the same time, social media is being used as a weapon to bring down or destroy somebody's reputation. Advocacy for freedom of speech has been a hot topic since ages. When an individual or a group of individuals is targeted, demeaned or abused by using foul language or any other hatred laden words, which affects their individuality or any other personal traits (behavioural, physical, mental, cultural or traditional) can be referred as 'abuse'. This abuse when encountered in online social media is known as 'online abuse'. Online abuse may also be depicted by online misrepresentation and cyberbullies [18]. It is believed that the reason for rising hate and abusive speech on the Internet could be because of the anonymity of users, possible to some extent by social media [20]. Thus people tend to display a more aggressive front where they take out their frustration and anger upon strangers on online platforms, and because of lack of repercussions, hate speech continues to prosper on social networking sites.

Although there have been attempts and studies to alleviate online abuse and hate speech [13] [8], yet online harassment permeates social media. To curb the prevailing trends of online hate speech and its other forms such as abusive speech, academicians and researchers are striving to develop a better way of identifying such content online. It is a crucial step in the right direction because social media now holds power to influence masses. Personal attacks are an increasing reason for mental stress [17]. Efficient detection of abusive content online might also aid in timely removal of such content from public websites [3]. Apparently, English has been the widely chosen language for research based on hate speech detection [19]. This paper focuses on abuse and hate detection in Indian languages. There are currently 23 official languages in India[2], however majority of them are still low resource or no resource when it comes to lexical resources [14]. Hence, abuse detection in this case becomes a tedious task. As some words in case of some languages may mean different in other languages and malicious groups or negative instinct people find it as a loophole to make their negative remarks and abuse in their local language.

It is in this light, this paper attempts to explore various methods to detect abusive content in Indian languages. Primarily, classical machine learning algorithms have been implemented with Indic embeddings, but new and efficient transformer based models prove to fit the data better. In this work, we perform the following:

- Experiments are performed to compare baseline classical machine learning model (Naive Bayes + Logistic Regression using indic tokenizer) with transformer based architecture (MuRIL Bert) whereby, MuRIL Bert outperformed Naive Bayes mode. This provides clear direction to further explore Transformer based models.
- Next, experiments are performed on both raw data and pre-processed data (transliterated and cleaned) using four state-of-the-art transformer-based models vis à vis XLM-RoBERTa, indic-BERT, MurilBert and mBERT are used whereby, XLM Roberta with BiGRU outperforms other models.
- Thereafter, the experimental setup of the best performing model XLM-RoBERTa is fed with emoji embeddings (see Section 3.4 and 4), that leads to further enhancement of overall performance of the proposed abuse detection methodology.
- The above experimental setup (XLM-RoBERTa with emoji embeddings) is trained with the combined dataset of 13 Indic languages. Next, the performance of this combined model is compared with four individual language models. The performance of combined model surpassed those of the individual models in terms

[2] https://www.newworldencyclopedia.org/entry/Languages_of_India

of F1 score and accuracy, supporting the fact that combined model fits the data better possibly due to its code-mixed nature.

The rest of the paper is organized as follows: Section 2 presents the related studies conducted in the recent past. Section 3 talks about the dataset used and its distribution. Section 4 lays down the proposed methodology. Section 5 describes the experimental setup and discusses the results in detail. Section 6 concludes the paper.

2 RELATED WORKS

According to an health affairs online hatred and abusive behaviour which can be either in form of aggressiveness, offensiveness or irony [23] impacts overall physical and mental health [17]. Studies performed using online social media often explores Twitter [16], [6], [9], [15], however, some other studies have also explored other platforms such as Facebook [1], [10], Reddit [12], 4chan, and 8chan [22]. These works have collected data using specific keywords or hashtags [16], or analysing specific pages [1], communities [22] which foster hate and abusive speech.

The works related to hate and abusive speech have been studied from many different aspects such as analysis from the perspective of religion [27], [1], sexism and misogyny [12], sexual minorities [16], migrants and refugees [6], [23], during elections [15], to name a few. For example, in [27] authors presented an approach for classifying antisemitism approach by focusing on common "dirty word" which could be evaded normally. In [1], based on posts from social media, a typology for characterizing anti-Muslim hate is put forward. In [12] researchers explored specific Reddit community called *manoshpere* which contains hate speech against females. Similarly, in [6] using Twitter data analysis was performed to study hate speech for migrants and refugees. The study reported that discussion was more against migrants compared to refugees. During German's 2017 election, hate speech was studied in political settings [15].

At abstract level, works related to hate and abusive speech can be divided into two categories, namely i) descriptive and ii) predictive. Descriptive works related to hate speech involves analysing hate comments [10], creation of taxonomies [1], spreading of hate speech [12], relationship between online hate speech and mental health [17], analysing various types of hate speech in online political communities [22], etc. In comparison, the works related to prediction, proposes various techniques for detecting hate speech [26], [19], and data annotation issues [28]. Predictive works generally include usage of user features and textual features for hate speech detection [26], [19]. Recently proposed approaches for detection of hate speech includes deep neural network based solutions, for example combining convolutional and gated recurrent networks [29], Recurrent Neural Network (RNN) classifiers [19] , and state of the art transformer's based solutions [11]. In another work, authors show that performance of hate speech detection improves by using multiple hate terms lists. [7]

Most of the works related to hate and abusive speech have explored English data. However, works have also used other languages such as Italian [10], Spanish [6], and German [15]. There have been works in less resource languages such as Danish, Greek, Turkish [2]. With respect to Indian languages, which we have also explored in this work, most of the works have mostly explore two to three languages (English, Hindi and Marathi) [2], [24], [11], [4]. The work closest to our work is [21], where authors presented multilingual offensive language identification with transformers. They also propose XLM- RoBERTa as the base transformer model but only tested for six languages, namely Bengali, Hindi, Kannada, Malayalam, Tamil, Urdu and English. In comparison, our work presents a mixed-model which has been trained on thirteen Indian languages for predicting abusive speech detection.

Language	Total Instances	Abusive	Non-abusive
Hindi	307180	153747 (50%)	153433 (50%)
Telugu	97012	48551 (50%)	48461 (50%)
Marathi	72044	27367 (34%)	44677 (66%)
Tamil	69497	34705 (50%)	34792 (50%)
Malayalam	40965	9216 (22.5%)	31749 (77.5%)
Bengali	22835	11407 (50%)	11428 (50%)
Kannada	13943	6989 (50%)	6954 (50%)
Odia	10974	5499 (50%)	5475 (50%)
Gujarati	8828	4402(50%)	4426(50%)
Haryanvi	8812	4417 (50%)	4395 (50%)
Bhojpuri	5804	2887 (50%)	2917 (50%)
Rajasthani	4368	2185 (50%)	2183 (50%)
Assamese	2780	1284 (47%)	1496 (53%)

Table 1. Language-wise distribution of abusive as well as non-abusive instances

3 DATASET

This section describes the dataset used for the experiments. Dataset is provided in a Kaggle competition for the Abuse Detection Challenge Dataset[3]. We thank ShareChat[4] for providing this dataset. This dataset consists of 13 low resource Indic languages, which is as an umbrella term comprising of languages that belong to the Indo-Aryan branch of the Indo-European languages. Dataset consists of various features including language, post index, commentText, report count comment, report count post, like count comment, like count post and label. The assumption that an abusive comment will have high report count on the comment didn't stand true for this dataset as the the distribution showcased that the majority data had 0 value. Hence, the focus of the work was based on 3 columns : language, commentText and label. The feature commentText refers to the text or the main content of the comment which is used to train the models, language refers to its annotated language which is used for comparative study whereas label is the column with ground truth. Presence of 0 as a label shows that the comment is not abusive and 1 shows that it is abusive. The language column of the dataset has 13 unique values, namely Hindi, Telugu, Marathi, Bengali, Rajasthani, Malayalam, Odia, Tamil, Gujarati, Kannada, Assamese, Haryanvi and Bhojpuri. The distribution of instances for various languages are different. The Table 1 contains instances of abusive as well as non-abusive available per language. This paper provides the percentage of both the classes with respect to the total instances to provide an information if the dataset for a particular language is balanced or not. Table 1 shows language-wise distribution of abusive as well as non-abusive instances.

3.1 Dataset Experimental Modifications

This section describes the preparation of the dataset before feeding it to the classification algorithms. There is a lack of tools for processing Indic languages, especially when the data is code-mixed. This dataset comprises of sentences weakly identified with 13 Indic languages. The inter class distribution is approximately balanced. Data pre-processing has been done in three steps which correspond to the three sub-sections 3.1, 3.2 and 3.3. Figure 1 depicts data pre-processing example of mixed script case of Hindi and English languages.

[3] https://www.kaggle.com/c/iiitd-abuse-detection-challenge
[4] https://sharechat.com/

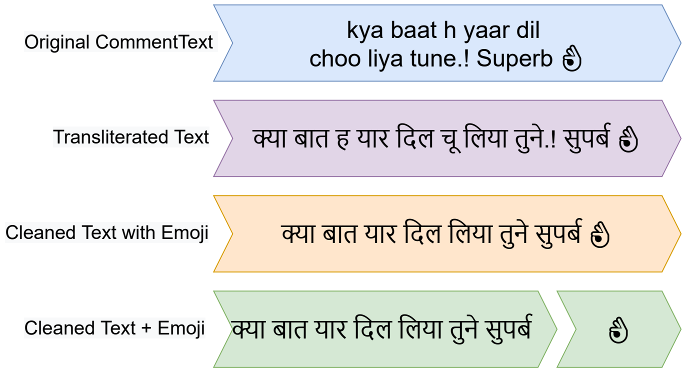
Original CommentText
kya baat h yaar dil choo liya tune.! Superb 👌
Transliterated Text
क्या बात ह यार दिल चू लिया तुने.! सुपर्ब 👌
Cleaned Text with Emoji
क्या बात यार दिल लिया तुने सुपर्ब 👌
Cleaned Text + Emoji
क्या बात यार दिल लिया तुने सुपर्ब
👌

Fig. 1. Data preprocessing : Sentence in English can be translated to "what's the matter dude, you touched my heart.! Superb"

3.2 Transliteration

Online written indic language based texts are combination of two types of data. One which is code mixed, that means the actual language of the text is different from which it is written in, usually English for example, 'Original Comment Text' in Figure 1. Second is the one where text is written in the native script of the language of the text. As the given data is written in a mix of English and native script of the language of the user, the training and the testing data were transliterated to the original native script of the language the sentence is identified as. Code-mixed data makes it incompatible for processes like stopword removal, automatic translation etc. Scarce number of open source libraries like indic nlp and inltk are available for handling and preprocessing multilingual data. In this case, indic nlp [5] library is used for transliterating employed dataset into the language of the script. A simple example of transliterated text from the library with correctness can be seen in Figure 1.

In this instance, the original text was in Hindi language with a script written in English. But after transliteration, the text is mapped into the Hindi script with substantial accuracy.

3.3 Data Cleaning

One of the most important part of pre-processing of data is to refine it by removing unwanted words, sentences or other parts of string irrelevant to the task. There were multiple steps taken to extract useful data from raw sentences. Stopword removal is a crucial step to remove unnecessary extra words. There are very few resources with respect to this step available but stopwords are collected together for individual languages from multiple sources. Table 2 shows the number of stopword databases created per language. The stopwords were collected from online available libraries and GitHub sources. Due to unavailability, there are some languages on which individual work could not be done as stop words were not available due to being a low resource language. These languages include Kannada, Rajasthani, Malayalam, Punjabi, Bhojpuri and Assamese. Python library named advertools [6] was used as the primary source of stopwords. Data contains various irrelevant information which is of no use to model, like, Hashtags, Tags and URLs. Hence these parts of sentences along with other regular expressions and punctuation were also removed from the data. As it can see in Figure 1, few common hindi words and exclamation mark has been removed from the transliterated text.

[5] https://github.com/anoopkunchukuttan/indic_nlp_library
[6] https://github.com/eliasdabbas/advertools

Language	Number of Stopwords
Hindi	914
Marathi	198
Bengali	556
Tamil	125
Telugu	46
Odia	68
Gujarati	1232

Table 2. Language-wise instances of abusive as well as non-abusive available

3.4 Emoji Data Processing

Emojis play a huge role in conveying emotions online. Hence, they can play a vital role in determining the sentiment or intent of the text. The emoji was separated from the text using demoji library [7] as it can be seen in the last tab of Figure 1. Then the embeddings of these emojis were extracted in numerical array format using emoji2vec library [8] along with gensim library [9]

4 METHODOLOGY

The proposed methodology primarily based on XLM-RoBERTa transformer based model along with BiGRU layer and 3-step data processing as mentioned in Section 3. Figure 2 depicts the model architecture used to differentiate between abusive and non-abusive content.

[7] https://pypi.org/project/demoji/

[8] https://github.com/uclnlp/emoji2vec

[9] https://radimrehurek.com/gensim/auto_examples/tutorials/run_word2vec.html

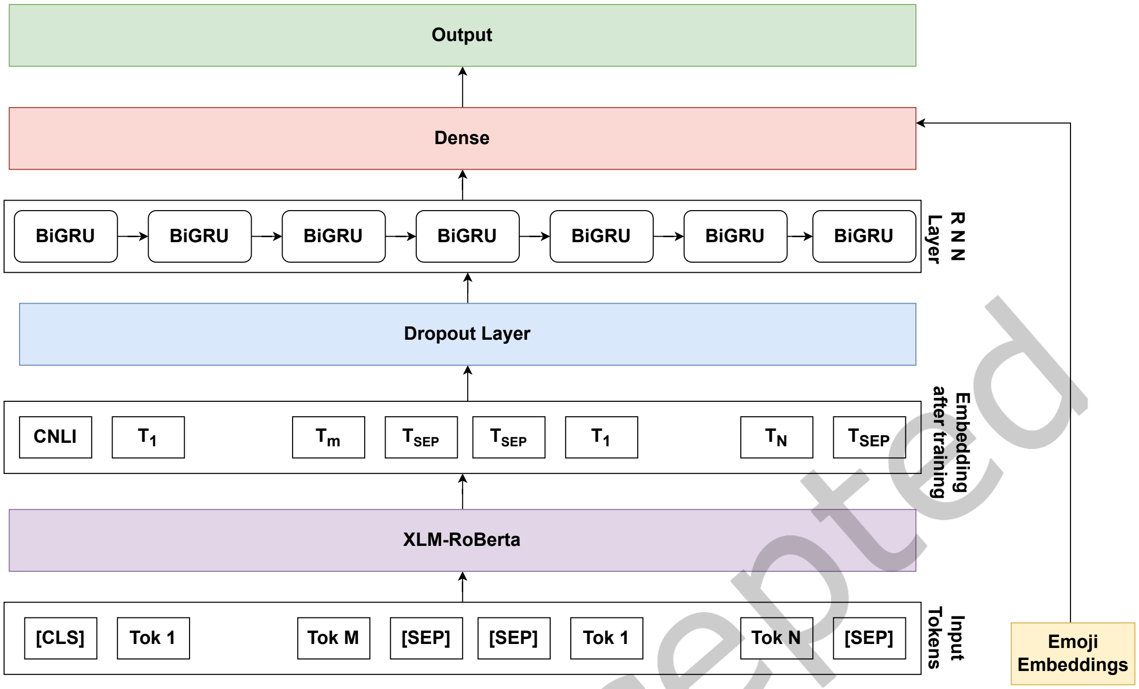

Fig. 2. Proposed Architecture

4.1 XLM RoBERTa

XLM RoBERTa is a multilingual model that has been trained in over 100 languages. The greatest difference between XLM RoBERTa and the original is the data used to train the models, its quality and quantity. It is very scalable especially obvious in languages with limited resources and hence it helps and best suits to the employed dataset with 13 languages all of them are more or less low resource in the NLP space. Its training method mimics a single language RoBERTa model, in that the only training aim is a masked language model.

4.2 BiGRU

BiGRU (Bidirectional Gated Recurrent Unit) is a recurrent neural network that is unique. It combines the gates of oblivion into a single entity. It combines hidden and cell states while converting the input gate to a single update gate. The BiGRU layer's main goal is to extract the text's deep characteristics from the input text vector. The relationship between contexts can be learned more extensively and semantic coding can be performed after feature extraction from the BiGRU layer.

Dropout was added between BiGRU layer and fully connected dense layer to overcome the issue of overfitting. To enhance the performance of the model, Emoji embeddings were concatenated with the outputs from the BiGRU layer before feeding them into the dense layer. This dense layer is further connected to the softmax layer which is utilised to provide the output.

4.3 Three-staged approach

This paper follows a three-level comparison approach to classify the data and find the best possible solution in terms of data and model used. Experiments for all these three stages are mentioned in detail in Section 5

- Stage 1 - Native Model Comparison : Classical ML algorithm like Naive Bayes + Logistic Regression using Indic tokenizer [10] is compared with one transformer based model : MurilBert to give starting direction to the work.
- Stage 2 - Transformer Based Model Comparison with DataProcessing based fine-tuning : Four transformer based models are compared - mBert, Muril Bert, Indic Bert and XLM Roberta. The effect of data cleaning and transliteration has been observed for the above models and it is observed what difference does emoji based embeddings bring to the table.
- Stage 3 - Individual Language models v/s single model : Herein, the objective is to compare the best performing method in Stage 2 with individual language based models for 4 languages, Hindi, Telugu, Marathi and Tamil to analyse the robustness of a single model.

5 EXPERIMENTAL SETUP AND RESULTS

5.1 Experimental Setup

All these experiments were performed with a batch size of 256, lr of 1e-6 and maxlen of 128. Tensorflow was used to train the model. The models were trained using a TPU provided by google colab with 25 GB of RAM. A seed of 42 was set to obtain reproducible results. The training dataset was split into 2 subsets: the training set and validation set. This split was performed in a 90 : 10 manner. This validation set helps to prevent overfitting of the model. A callback with patience of 3 was used to train the model to prevent overfitting. This check was performed on the 'validation loss' feature during the training. Refer to Figure 1 and Figure 2 for experimental pipeline.

5.2 Results and Analysis

5.2.1 Comparison with Baseline Models . First simple techniques were employed to obtain a baseline performance. The data was tokenized using IndicNLP tokenizer and given as input into a naive bayes plus logistic regression model to obtain baseline performance before applying complex deep learning algorithms. Then its performance was compared with a MuRIL Bert model which is a Bert-based and pre-trained on data from indic resources which contain most of the languages in the dataset provided.

The Naive Bayes + Logistic regression model renders a F1 score of 86.3% on the test set. This provided a hint that the transformer based model will outperform the conventional machine learning models in this dataset. Hence, the setup proceeds with the multilingual Bert-based models on the commentText feature. Table 3 reports the F1 scores on test dataset for baseline models.

Model	F1 score
Naive Bayes + Logistic Regression using indic tokenizer	0.863
MuRIL Bert	**0.875**

Table 3. F1 scores : Baseline Model Comparison on Test Set

After MuRIL Bert performs better then various experiments were performed with various transformer based architecture to obtain maximum performance. These experiments were performed with comment Text (raw data) as well as transliterated cleaned text.

5.2.2 Transformer Based Model Comparison. mBert performed with 86.51% accuracy, 86.78% precision, 84.03% recall, 93.07% AUC and 84.93% F1 score. Using a threshold value of 0.6 due to decreased value of recall as compared to precision resulted in a F1 score of 87.2 on the test set. MuRIL Bert, XLM RoBERTa, as well as indic

[10] https://github.com/ltrc/indic-tokenizer

Bert models experiment with the comment Text feature. MuRIL Bert obtained values of 87.24%, 87.87%, 84.45%, 93.45% and 85.68% of accuracy, precision, recall, AUC and F1 score. XLM RoBERTa is especially designed for low resource indic languages, and hence obtain values of 87.27%, 86.92%, 85.78%, 93.06% and 85.94% of accuracy, precision, recall, AUC and F1 score. XLM RoBERTa outperforms all other models so far with a performance of F1 score of 87.603. These models were then run on the transliterated commentText feature to compare them with non-transliterated features. In this XLM RoBERTa performance increased to 87.29%, 86.38%, 86.54%, 93.7% and 86.05% values of accuracy, precision, recall, AUC and F1 score. This model outperformed all other models with a value of 87.869. After this performance, another model was trained by concatenating the embedding of emoji extracted using a pre-trained embedding into the embedding obtained from the transformer-based model. Here cT means commentText data whereas TcT means transliterated cleaned commentText data. Table 4 reports the comparative values on the various evaluation metrics for the validation dataset. Figure 3 presents a visual summary of the results reported in. It can be observed from this figure XLM RoBERTa with Transliterated text and emoji embeddings outperforms other methodologies.

Model	Val Loss	Val Accuracy	Val Precision	Val Recall	Val AUC	Val F1 Score
mBert with cT	0.332	0.865	0.867	0.840	0.930	0.8495
MuRIL Bert with cT	0.322	0.872	0.878	0.844	0.934	0.856
XLM RoBERTa with cT	0.344	0.872	0.869	0.857	0.930	0.859
Indic Bert with cT	0.358	0.851	0.855	0.821	0.918	0.833
mBert with TcT	0.327	0.867	0.877	0.833	0.932	0.850
MuRIL Bert with TcT	0.320	0.872	0.877	0.846	0.934	0.857
XLM RoBERTa with TcT	0.345	0.872	0.863	0.865	0.937	0.860
Indic Bert with TcT	0.374	0.842	0.847	0.809	0.910	0.823
XLM RoBERTa with TcT + emoji embedding	**0.313**	**0.877**	**0.877**	**0.858**	**0.940**	**0.864**

Table 4. Validation metric comparison

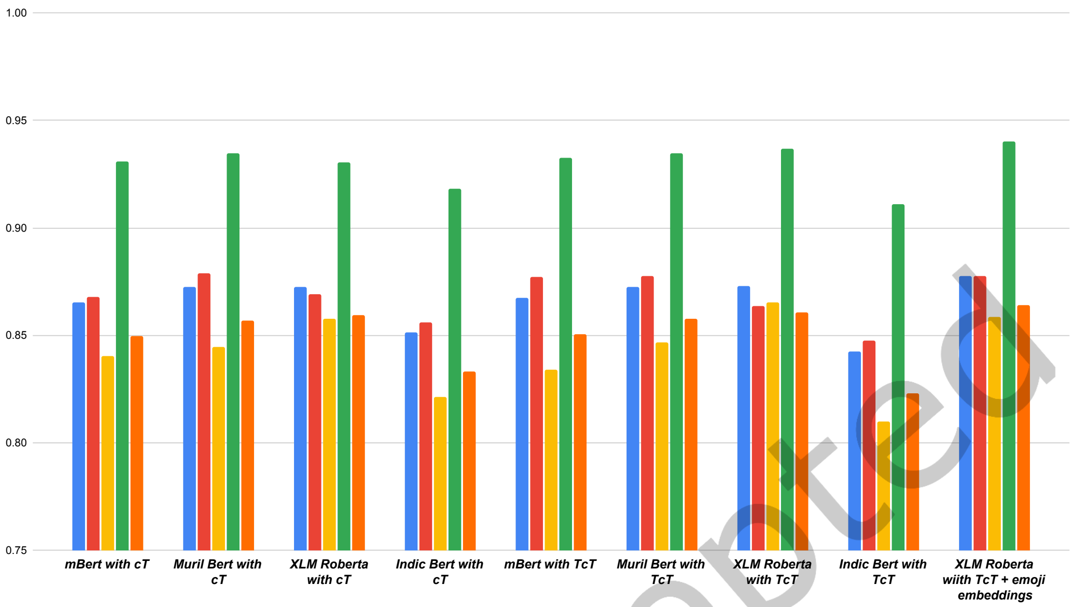

Fig. 3. Metric Evaluation on Validation set

5.2.3 Language-specific model versus Generic Model. After choosing the methodology from the best scoring model, it is compared with language specific models. Table 5 presents the results of language specific model (Hindi[11], Telugu[12], Tamil[13] and Marathi[14]) as compared to generic model (XLM-RoBERTa). Experiments for comparison based study were performed for these four languages due to unavailability of pre-trained transformer based models for other Indic languages.

Language	Individual F1	XLM-RoBERTa F1	Individual accuracy	XLM-RoBERTa accuracy
Hindi	0.87	0.91	0.87	0.91
Telugu	0.85	0.88	0.86	0.88
Tamil	0.92	0.93	0.92	0.93
Marathi	0.60	0.80	0.69	0.80

Table 5. Results of Language-specific model v/s Generic Model

As it can be observed in Table1 out of all the languages, Marathi and Malayalam have observable differences in number of abusive and non abusive instances. Intra-class balancing was tried for these two languages by augmenting instances belonging to the minor class. However, no significant changes could be reported for this experiment.

[11] https://huggingface.co/flax-community/roberta-hindi

[12] https://huggingface.co/neuralspace-reverie/indic-transformers-te-roberta

[13] https://huggingface.co/abinayam/gpt-2-tamil

[14] https://huggingface.co/DarshanDeshpande/marathi-distilbert

6 CONCLUSION

This paper performs experimental analysis to detect abuse in online social media. For this purpose, a dataset of 13 Indic code-mixed languages is used. Further, the paper lays down an extensive experimental study with Native ML models, state-of-the-art transformer based models such as XLM-RoBERTa, Indic-BERT, Muril Bert and mBERT to show that XLM Roberta outperformed others. This paper further demonstrates that this performance can be improved by suitable data pre-processing and leveraging emoji based data embeddings to obtain maximum performance. This method performs decently but could refrain from performing exceptionally well due to the presence of sarcastic text in the text available. This model obtains a F1 score of 88.096 on the test set. This paper also highlights how single multilingual model outperforms individual language models.

However, due to the unavailability of pre-trained model resources for many languages (owing to the low resource nature of Indic languages), the above mentioned can't be exclusively claimed for each language but a general view of the performance comparison can be clearly observed. The reason behind such behavior could be seen:

(1) When the dataset was looked at manually it was found that the annotation of few instances, in terms of their language label, was not appropriate, e.g. the sentence is written and meant both in English but the label of language was Telugu. This creates a hindrance in individual language models and its performance.
(2) Majority of sentences are code-mixed and contain words from multiple languages. As the source of the dataset is a social media app where the user can feed in multiple languages written in different scripts, a single combined model allows a better learning space for the input data rather than individual language models.

There can be multiple future directions to extend this work. In the first direction, the dataset could be scaled up by making it an integrated multi-sourced dataset. Additionally, novel and efficient models can be devised for handling other low resource languages. Furthermore, having different pre-processing toolkits for individual languages might significantly improve the quality of input being fed into the transformer based models.

REFERENCES

[1] Imran Awan. 2016. Islamophobia on Social Media: A Qualitative Analysis of the Facebook's Walls of Hate. *International Journal of Cyber Criminology* 10, 1 (2016).

[2] Somnath Banerjee, Maulindu Sarkar, Nancy Agrawal, Punyajoy Saha, and Mithun Das. 2021. Exploring Transformer Based Models to Identify Hate Speech and Offensive Content in English and Indo-Aryan Languages. *arXiv preprint arXiv:2111.13974* (2021).

[3] James Banks. 2010. Regulating hate speech online. *International Review of Law, Computers & Technology* 24, 3 (2010), 233–239.

[4] Mehar Bhatia, Tenzin Singhay Bhotia, Akshat Agarwal, Prakash Ramesh, Shubham Gupta, Kumar Shridhar, Felix Laumann, and Ayushman Dash. 2021. One to rule them all: Towards Joint Indic Language Hate Speech Detection. *arXiv preprint arXiv:2109.13711* (2021).

[5] Grant Blank and Christoph Lutz. 2018. Benefits and harms from Internet use: A differentiated analysis of Great Britain. *New Media & Society* 20 (02 2018), 618–640. https://doi.org/10.1177/1461444816667135

[6] Carlos Arcila Calderón, David Blanco-Herrero, and María Belén Valdez Apolo. 2020. Rejection and hate speech in Twitter: Content analysis of Tweets about migrants and refugees in Spanish. *Revista Española de Investigaciones Sociológicas (REIS)* 172, 172 (2020), 21–56.

[7] Animesh Chaturvedi and Rajesh Sharma. 2022. minOffense: Inter-Agreement Hate Terms for Stable Rules, Concepts, Transitivities, and Lattices. In *Proceedings of the 9th IEEE International Conference on Data Science and Advanced Analytics*.

[8] Raphael Cohen-Almagor. 2011. Fighting hate and bigotry on the Internet. *Policy & Internet* 3, 3 (2011), 1–26.

[9] João Guilherme Routar de Sousa. 2019. Feature extraction and selection for automatic hate speech detection on twitter. (2019).

[10] Fabio Del Vigna12, Andrea Cimino23, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*. 86–95.

[11] Zaki Mustafa Farooqi, Sreyan Ghosh, and Rajiv Ratn Shah. 2021. Leveraging Transformers for Hate Speech Detection in Conversational Code-Mixed Tweets. *arXiv preprint arXiv:2112.09986* (2021).

[12] Tracie Farrell, Miriam Fernandez, Jakub Novotny, and Harith Alani. 2019. Exploring misogyny across the manosphere in reddit. In *Proceedings of the 10th ACM Conference on Web Science*. 87–96.

[13] Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. 2015. *Countering online hate speech*. Unesco Publishing.

[14] Vedika Gupta, Nikita Jain, Shubham Shubham, Agam Madan, Ankit Chaudhary, and Qin Xin. 2021. Toward Integrated CNN-based Sentiment Analysis of Tweets for Scarce-resource Language—Hindi. *Transactions on Asian and Low-Resource Language Information Processing* 20, 5 (2021), 1–23.

[15] Sylvia Jaki and Tom De Smedt. 2019. Right-wing German hate speech on Twitter: Analysis and automatic detection. *arXiv preprint arXiv:1910.07518* (2019).

[16] Vittorio Lingiardi, Nicola Carone, Giovanni Semeraro, Cataldo Musto, Marilisa D'Amico, and Silvia Brena. 2020. Mapping Twitter hate speech towards social and sexual minorities: a lexicon-based approach to semantic content analysis. *Behaviour & Information Technology* 39, 7 (2020), 711–721.

[17] Shruthi Mohan, Apala Guha, Michael Harris, Fred Popowich, Ashley Schuster, and Chris Priebe. 2017. The impact of toxic language on the health of reddit communities. In *Canadian Conference on Artificial Intelligence*. Springer, 51–56.

[18] Justin W Patchin and Sameer Hinduja. 2006. Bullies move beyond the schoolyard: A preliminary look at cyberbullying. *Youth violence and juvenile justice* 4, 2 (2006), 148–169.

[19] Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Effective hate-speech detection in Twitter data using recurrent neural networks. *Applied Intelligence* 48, 12 (2018), 4730–4742.

[20] Harrison Rainie, Janna Quitney Anderson, and Jonathan Albright. 2017. *The future of free speech, trolls, anonymity and fake news online.* Pew Research Center Washington, DC.

[21] Tharindu Ranasinghe and Marcos Zampieri. 2021. An evaluation of multilingual offensive language identification methods for the languages of india. *Information* 12, 8 (2021), 306.

[22] Diana Rieger, Anna Sophie Kümpel, Maximilian Wich, Toni Kiening, and Georg Groh. 2021. Assessing the Extent and Types of Hate Speech in Fringe Communities: A Case Study of Alt-Right Communities on 8chan, 4chan, and Reddit. *Social Media+ Society* 7, 4 (2021), 20563051211052906.

[23] Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An italian twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*

[24] Arushi Sharma, Anubha Kabra, and Minni Jain. 2022. Ceasing hate with MoH: Hate Speech Detection in Hindi–English code-switched language. *Information Processing & Management* 59, 1 (2022), 102760.

[25] Michael Simonson. 2017. Social Media and Online Learning: Pros and Cons. *Distance Learning* 14, 4 (2017), 72–71.

[26] Elise Fehn Unsvåg and Björn Gambäck. 2018. The effects of user features on twitter hate speech detection. In *Proceedings of the 2nd workshop on abusive language online (ALW2).* 75–85.

[27] William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media.* 19–26.

[28] Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science.* 138–142.

[29] Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference.* Springer, 745–760.