EAI Endorsed Transactions
on Scalable Information Systems

Research Article EAI.EU

Truculent Post Analysis for Hindi Text

Mitali Agarwal[1], Poorvi Sahu[1], Nisha Singh[1], Jasleen[1], Puneet Sinha[2], Rahul Kumar Singh*[1]

[1]SoCS, University of Petroleum and Energy Studies, Dehradun, Uttarakhand 248001, India
[2]Bajaj Finserv, Pune, Maharashtra, India

Abstract

INTRODUCTION: With the rise of social media platforms, the prevalence of truculent posts has become a major concern. These posts, which exhibit anger, aggression, or rudeness, not only foster a hostile environment but also have the potential to stir up harm and violence.
OBJECTIVES: It is essential to create efficient algorithms for detecting virulent posts so that they can recognise and delete such content from social media sites automatically. In order to improve accuracy and efficiency, this study evaluates the state-of-the-art in truculent post detection techniques and suggests a unique method that combines deep learning and natural language processing. The major goal of the proposed methodology is to successfully regulate hostile social media posts by keeping an eye on them.
METHODS: In order to effectively identify the class labels and create a deep-learning method, we concentrated on comprehending the negation words, sarcasm, and irony using the LSTM model. We used multilingual BERT to produce precise word embedding and deliver semantic data. The phrases were also thoroughly tokenized, taking into consideration the Hindi language, thanks to the assistance of the Indic NLP library.
RESULTS: The F1 scores for the various classes are given in the "Proposed approach" as follows: 84.22 for non-hostile, 49.26 for hostile, 68.69 for hatred, 49.81 for fake, and 39.92 for offensive
CONCLUSION: We focused on understanding the negation words, sarcasm and irony using the LSTM model, to classify the class labels accurately and build a deep-learning strategy.

Keywords: Truculent Post, Hindi language, Sentiment Analysis, BERT, LSTM, NLP

Received on 28 December 2023, accepted on 29 March 2024, published on 04 April 2024

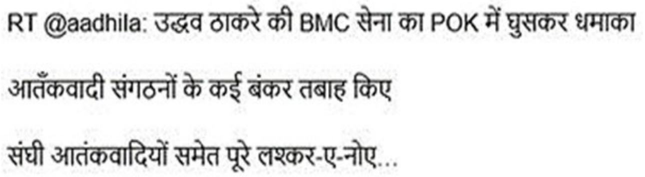
doi: 10.4108/eetsis.5641

*Corresponding author. Email: rahulcu25@gmail.com

1. Introduction

In recent years, there has been a greater emphasis on truculent word detection in Hindi language. This is due to the growing number of hate crimes and incidents of discrimination against minorities in India. The automatic detection of truculent words in text is a challenging problem that has received considerable attention in recent years. Since, COVID period, the majority of people are spending more time on social media, chat rooms, communication applications, and gaming servers, which has a negative impact on their mental health [1]. People use websites like Twitter, Facebook, WhatsApp, and other ones to express their views towards different communities, groups and topics related to education, entertainment and politics are being discussed here. The views may have positive or negative effects on the society depending on one's perspective. A variety of approaches have been proposed, but there is still no consensus on the best way to tackle the problem.
We review the existing literature on truculent word detection and propose a new approach based on deep learning. Firstly, we performed pre-processing to simplify the post then we have used multilingual BERT for word embedding and LSTM is used for classifying different class labels such as Hate, Fake, Defamation, Offensive and Non-hostile.

HATE: Posting negative responses towards certain groups of people, communities and religions which will target them in a specific way. Figure 1 shows the example of Hate class in textual data.
Example:



RT @aadhila: उद्धव ठाकरे की BMC सेना का POK में घुसकर धमाका

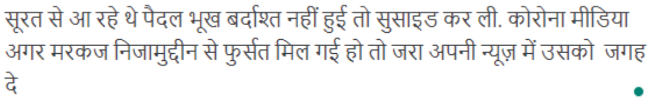
आतँकवादी संगठनों के कई बंकर तबाह किए

संघी आतंकवादियों समेत पूरे लश्कर-ए-नोए...

Figure 1: Hate class example

FAKE: Post which spreads misinformation or influences public opinion on certain topics. Figure 2 depicts the example of Fake class in textual data.

Example:

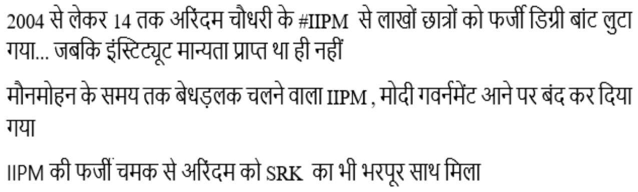


सूरत से आ रहे थे पैदल भूख बर्दाश्त नहीं हुई तो सुसाइड कर ली. कोरोना मीडिया अगर मरकज निजामुद्दीन से फुर्सत मिल गई हो तो जरा अपनी न्यूज़ में उसको जगह दे

Figure 2: Example of Fake class

DEFAMATION: Post to degrade the reputation of an organization or certain group of people. The example of defamation class is depicted in figure 3.

Example:



2004 से लेकर 14 तक अरिंदम चौधरी के #IIPM से लाखों छात्रों को फर्जी डिग्री बांट लुटा गया... जबकि इंस्टिट्यूट मान्यता प्राप्त था ही नहीं

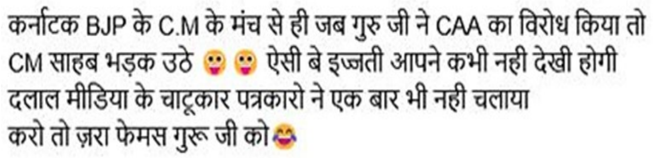
मौनमोहन के समय तक बेधड़लक चलने वाला IIPM , मोदी गवर्नमेंट आने पर बंद कर दिया गया

IIPM की फर्जी चमक से अरिंदम को SRK का भी भरपूर साथ मिला

Figure 3: Example of Defamation class

OFFENSIVE: Post which can be hurtful, disrespectful, insulting and damaging to the individuals. The example of offensive class is shown in figure 4.

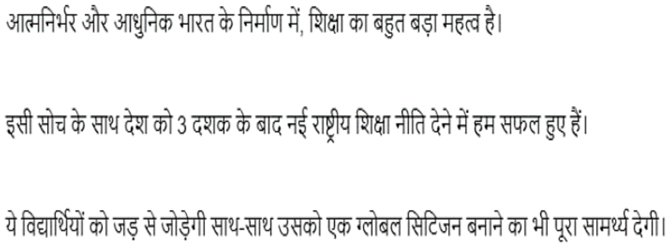
Example:



कर्नाटक BJP के C.M के मंच से ही जब गुरु जी ने CAA का विरोध किया तो
CM साहब भड़क उठे 😛😛 ऐसी बे इज्जती आपने कभी नही देखी होगी
दलाल मीडिया के चाटूकार पत्रकारो ने एक बार भी नही चलाया
करो तो ज़रा फेमस गुरू जी को😂

Figure 4: Example of Offensive class

NON-HOSTILE: A post which does not contain any hostility. Figure 5 depicts the example of Non-Hostile class in textual data.



आत्मनिर्भर और आधुनिक भारत के निर्माण में, शिक्षा का बहुत बड़ा महत्व है।

इसी सोच के साथ देश को 3 दशक के बाद नई राष्ट्रीय शिक्षा नीति देने में हम सफल हुए हैं।

ये विद्यार्थियों को जड़ से जोड़ेगी साथ-साथ उसको एक ग्लोबल सिटिजन बनाने का भी पूरा सामर्थ्य देगी।

Figure 5: Example of Non- Hostile class

Sentiment analysis is a powerful tool that can be used to detect hostile posts on social media and other online platforms. It is a form of Natural Language Processing that aims to recognise and interpret the sentiment expressed in text. It involves taking information from spoken or written language in order to understand how people feel and think about certain issues.

1.1. Problem Statement

Truculent post detection is a technology which is used to monitor the posts of social media users and put control on them. Its main objective is to improve online space, where users can speak their mind and create a healthy environment without negative posts. Truculent post detection detects that the user has not used negative language. It shows whether that user has posted a hostile post or not? With its help, we can block/delete these hostile messages and start managing them properly. In this process, we use AI and ML which analyses data and detects patterns so that we can block these unwanted messages.

The problem statement for truculent post detection in Hindi is the need for such a system that there is a great deal of hostility and negativity online which can lead to real world consequences. Also, current systems for detecting hostile words are not adequate and often fail to identify Hindi words. Finally, this project aims to develop a system that can accurately detect Hindi words with hostility and provide users with warnings accordingly.

Many models have been implemented to analyse the sentiments of the posts or text. Some of them are based on machine learning algorithms because they produce more accurate findings than conventional techniques like human coding or keyword-based approaches. The algorithm can find patterns in the data set that cannot be found manually, enabling it to better comprehend how people feel about specific subjects, goods, or services based just on their language usage. The flexibility of machine learning models also makes them the best choice for long-term sentiment tracking projects where accuracy must remain high throughout numerous iterations of a project's life cycle. Machine learning models are able to adjust themselves over time so that they continue to provide helpful insights even when new information becomes available. While machines excel at finding patterns within large amounts of data quickly, they often fail at understanding complex emotions such as sarcasm or irony since these concepts require higher-level cognitive processing that is beyond what current technology offers us today.

Nowadays, deep learning models are increasingly being used in this process due to their ability to capture complex patterns from large amounts of data. By leveraging the power of deep learning, businesses can gain deeper insights into customer sentiment than ever before. While deep learning algorithms are able to automatically find features without any prior knowledge of the dataset or labels connected with it, traditional approaches require manual feature engineering, which may be time-consuming and expensive. This enables more precise predictions by catching minute differences between various emotions that could otherwise be missed by humans or more straightforward machine-learning techniques like logistic regression, SVM and Naive Bayes classifiers.

To solve the problem of machine learning algorithms we have implemented the deep learning model LSTM due to its recurrent nature. LSTMs are able to effectively assess complicated associations between words over time, which is crucial when dealing with natural language processing tasks like sentiment analysis. Also, they have been demonstrated to perform better than conventional neural networks in textual datasets when it comes to capturing these kinds of associations since they have the capacity to learn from the past through back propagation through time (BPTT). Word embeddings can be effectively incorporated into LSTM networks, enhancing sentiment analysis performance even more. By using these word embeddings, LSTM models can capture the semantic meaning of words and their effect on sentiment analysis. LSTMs are also more flexible when learning complicated patterns from datasets because they use specialized memory cells within each LSTM unit, which allows them to capture and retain information over long periods, making them well-suited for modelling sequential data; as a result, they are better suited for analysing sentiments than common techniques like bag-of-words approaches or support vector machines (SVMs).

1.2 Contribution

The main contribution of the paper is as follows:

- In the original dataset, we have encoded labels of each post and performed pre-processing.
- We examine the existing literature on hostile word identification and provide a novel, deep learning-based strategy.
- For tokenization, we have used the Indic NLP library as it includes punctuations for the Indian language scripts such as the purna virama and the deergha virama. It also has the ability to accurately interpret complex text written in these languages.
- To generate accurate word embeddings for Hindi language we have used uncased multilingual BERT. Uncased mBERT offers a powerful method for representing words from any corpus as vectors, which may then be utilised as input features for different NLP applications.
- For truculent post detection we have used the LSTM model as it easily identifies complex patterns and relationships in language data which helps in understanding negation words like no, not etc., sarcasm and irony. Further, it becomes easy to classify class labels accurately.

2. Related Work

The detection of truculent posts on social media and other online platforms is a growing problem. As the use of these platforms increases, so does the potential for malicious actors to spread disinformation or incite violence. Researchers have created a number of methods for spotting aggressive comments in online interactions to solve this problem.

To solve the multi-level multi-class classification problem Kumar et al.[2] suggested a two level framework made up of statistical and BERT based classifiers. Jha et al.[3] presents a model that distinguishes offensive text from non-offensive text using a fastText-based model. Jayanthi et al.[4] used Task-Adaptive Pre-Training of Multilingual BERT models for Offensive Language Identification in Dravidian languages. Their model is a set of 6 various mBERT and XLMRoBERTa models, specially tuned for the job of detecting offensive words. Bhatnagar et al.[5] The task was separated between coarse and fine grained tasks. And investigated methodologies such as transformer and multitask learning before employing classic statistical classifiers such as SVM and polynomial logistic regression. Javier et al.[6] used a mixed methodological method for determining whether an aggressive tone is used during the campaign. They performed the quantitative and qualitative method for extracting informational insights. In this paper, Singh et al.[7] proposed different methods to detect fake reviews. They extracted aspects from the review and fed it into CNN for aspect replication learning and for fake reviews detection they passed the replicated aspect into the LSTM model. They demonstrated that deep neural networks outperform conventional approaches in complex computation. Schmidt et al.[8] carried out research on the automatic recognition of hate speech. They used Character-level methods for detection of hate speech, such as bags of words or embeddings as performed better than token-level approaches. To identify the toxic emotions, d'Sa et al.[9] used BERT and FastText embeddings for representation of words and deep learning techniques. They performed two approaches for binary and multi class classification, one of them is a method of extracting word embedding and then using DNN and another one is fine tuning of the BERT model which performed better. Roy et al.[10] used Task Adaptive pre training (TAPT) approach prior to fine-tuning of transformer-based architectures. Constructed a categorization architecture that takes emojis and segmented hashtags into account. Sachan et al.[11] concentrated on capturing the common features across domains and developed an aspect-based neural networks framework that incorporates information from the contexts and aspects of sentences. The main objective of the proposed model is to find domain-invariant traits based on aspect and sentence information in order to effectively transfer feelings across the domain.

In [12] authors classified hostile posts using One-vs-the-Rest neural network-based technique. To acquire the contextual representations of Hindi postings, they used multilingual BERT which was pre-trained also they used hybrid strategies for their studies. Their suggested model proved to be the most effective baseline model for identifying hostility in the Hindi messages, outperforming it producing F1 scores of 92.60%, 81.14%, 69.59%, 75.29%, and 73.01% for the categories "hostile," "fake," "hate," "offensive," and "defamation," respectively. Badjatiya et al.[13] classified tweets as racist, sexist or neither. For this they conducted extensive tests with variety of deep learning systems and concluded that deep neural networks and gradient-boosted decision trees outperform modern word/character n-gram approaches. To recognise racist and sexist remarks, Waseem et al.[14] provided a list of standards based on critical race theory which can be utilised to collect additional information and deal with the issue of a small but extremely active group of nasty people and discovered that a character n-gram based method offers a strong basis over word-length distribution. In [15] authors have introduced a transfer learning-based method for multi-dimensional hostile post identification in Hindi Devanagari script. In order to classify additional sub-tasks, this article uses attention-based pre-trained models that have been optimised on Hindi data and include the Hostile-Non hostile task as an auxiliary task. They have created a robust model that has the capacity to recognise biases and ambiguities that may have occurred during the collection or annotation of the dataset. In [16] the author presents the first labeled dataset on BanglaFake news, it implies that models based on neural networks cannot always outperform linear classifiers with conventional language data.

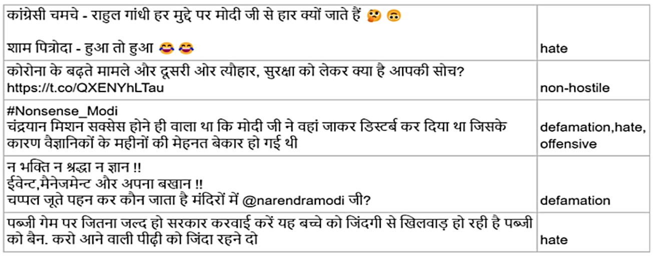
2.1. Problem Definition

In this section, we will discuss about the problem that we identified in existing approaches. Machine translation approach, translates the Hindi post to English then applies the analysis on English text which might change the meaning and emotion behind the post such as sarcasm and humour cannot be accurately translated into a language in which it was not initially written. Machine learning models are excellent at quickly identifying patterns in vast volumes of data but they frequently fall short when it comes to understanding complex emotions like sarcasm or irony since these concepts demand higher-level cognitive processing than what our present model can provide.

3. Dataset

Table 1 provides a summary of the dataset's statistics [1]. Train, Test, and Validation dataset segments have been for training, validation, and testing. This dataset is then pre-processed using a natural language processing toolkit i.e nltk.

created. There are a total 8192 posts which are divided into two categories Hostile and Non-hostile. 4358 Examples fit into the area of non-hostility out of 8192 posts and remaining 3834 posts belong to the hostile category. The hostile category is further divided into four sub-categories: *fake, hate, offensive and defamation*. There are 1638, 1132, 1071 and 810 posts for fake, hate, offensive and defamation respectively. One post may fall under one or more of the hostile class categories depicted in figure 6.



कांग्रेसी चमचे - राहुल गांधी हर मुद्दे पर मोदी जी से हार क्यों जाते हैं 🤔 🙃 शाम पित्रोदा - हुआ तो हुआ 😂 😂	hate
कोरोना के बढ़ते मामले और दूसरी ओर त्यौहार, सुरक्षा को लेकर क्या है आपकी सोच? https://t.co/QXENYhLTau	non-hostile
#Nonsense_Modi चंद्रयान मिशन सक्सेस होने ही वाला था कि मोदी जी ने वहां जाकर डिस्टर्ब कर दिया था जिसके कारण वैज्ञानिकों के महीनों की मेहनत बेकार हो गई थी	defamation,hate, offensive
न भक्ति न श्रद्धा न ज्ञान !! ईवेन्ट,मैनेजमेन्ट और अपना बखान !! चप्पल जूते पहन कर कौन जाता है मंदिरों में @narendramodi जी?	defamation
पब्जी गेम पर जितना जल्द हो सरकार करवाई करें यह बच्चे को जिंदगी से खिलवाड़ हो रही है पब्जी को बैन. करो आने वाली पीढ़ी को जिंदा रहने दो	hate

Figure 6: Different categories of dataset

Dataset provides valuable insights into the nature of online hostile speech and its prevalence in Hindi language. Recent studies have shown that these datasets can be used to develop effective machine learning models for detecting and classifying hateful posts on social media platforms. Statistics from these datasets indicate that there is a significant amount of hate speech present in Hindi language content on the internet, with more than one-third of all posts containing some form of offensive or derogatory language. This highlights the importance for researchers to continue developing better methods for recognizing such posts and understanding their implications across different languages. The word cloud for the training, testing, and validation dataset is displayed in figure 7.

Figure 7: Word Cloud of dataset (a. Training, b. Testing, c. validation)

4. Proposed Architecture

Figure 8 shows the proposed architecture of this research work. Twitter postings are collected from many users and pooled into one dataset, which is then split into three sections.

Table 1: Statistics of dataset

	Hostile					Non-Hostile
	Fake	Hate	Offensive	Defamation	Total	
Train	1144	792	742	564	2678	3050
Test	334	237	219	169	780	873
Validation	160	103	110	77	376	435
Total	1638	1132	1071	810	3834	4358

The pre-processing steps include:

- Removing the punctuations from the text.
- Removing stopwords: We have compiled a list of Hindi stop words that we have then eliminated from the text.
- Removing all the URLs, Hashtags, Emojis, Mentions and other special characters from the text to simplify it.
- Then the text is tokenized using the indic nlp tokenizer library, which is more efficient in tokenizing Indian languages such as Hindi in the comparison with the tokenize library provided by the nltk.

The texts (or the posts) now have been tokenized. Each sentence/post is now a list of words or tokens.

$$S = \{ w_1, w_2, w_3, \; . \; . \; . \; . \; . \; . \; . \; , w_n\}$$

where S is the sentence and w1, w2, w3, ... , wn refers to words. This list of tokens is then further provided to perform the embedding. Here, we have used the flair library.

The flair library can be used to combine different word and document embedding. Here, we have used and performed the bert-base-multilingual-uncased embedding.

$$BERT\ (w_i) = \sum_{k=1}^{K} \alpha_k \frac{\sum_j^p g_j^k}{p} \qquad (1)$$

where the kth layer's weight coefficient is αk, byte pair encoding token (g_j^k) is the jth hidden state of kth layer, where, 1 ≤ k ≤ K, 1 ≤ j ≤ p. To obtain the sentence embedding vectors (es) for each word (wi), we find the mean of embedding of the words.

BERT is a transformers model pre-trained on a large corpus of multilingual data. It has 12 bidirectional self-attention heads and 768 hidden units in a 12-layer encoder stack. It is a 12- layer of encoder stack with 12 bidirectional self-attention heads and 768 hidden units. It consists of 12-heads and 110M parameters, and can be performed on 102 different languages. After we have the contextual embeddings with the help of BERT, the data is converted into the numerical form sentences. Then these sentences are given as input to the Bidirectional LSTM model (Long-Short Term Memory). The LSTM networks are ideal for processing and predicting sequential data. The LSTM model is used to predict the class labels i.e., if a post is non-hostile or hostile.

If the post is labeled as hostile then it will predict further, which category of hostile post does it belongs to:

- Hate
- Defamation
- Fake
- Offensive

After the classification of the class label, the F1 score is calculated, as it is a more reliable confusion metric.

5. Experimental Analysis

In experimental analysis, for refining data pre-processing of posts is done using various techniques such as removal of stopwords in hindi and english, removal of punctuations, removal of emoticons, removal of regular expressions, tokenization and word embedding.

The subsequent subsection describes the experimental setup, result and analysis and discussion.

5.1 Experimental Setup

Pandas, Numpy, Keras, and Pytorch are used to implement deep learning models. For preprocessing re is used for removal of URLs, emoticons and special characters, indic library is used for tokenization of the hindi post and corpus of stopwords and punctuations are used. Flair.embeddings library is used to import multilingual Bert for word embedding. LSTM library is imported from Keras for classifying the posts. To create the accuracy report of the model sklearn.metrics library is used. The overall implementation is performed on Google Colab GPU.

5.2 Result & Analysis

In table 2, we compared all the approaches used for the classification of hostile and non-hostile posts. For our proposed approach i.e. LSTM we got F1 score, Non-hostile(84.22%), hate(49.26%), fake(68.69%), offensive(49.81%) and defamation(39.92%).Among all the approaches which make use of multilingual BERT, LSTM performed better but XML-roberta based BERT outperforms mBERT in all cases. So,if we use LSTM with XML-roberta based BERT the accuracy might increase.

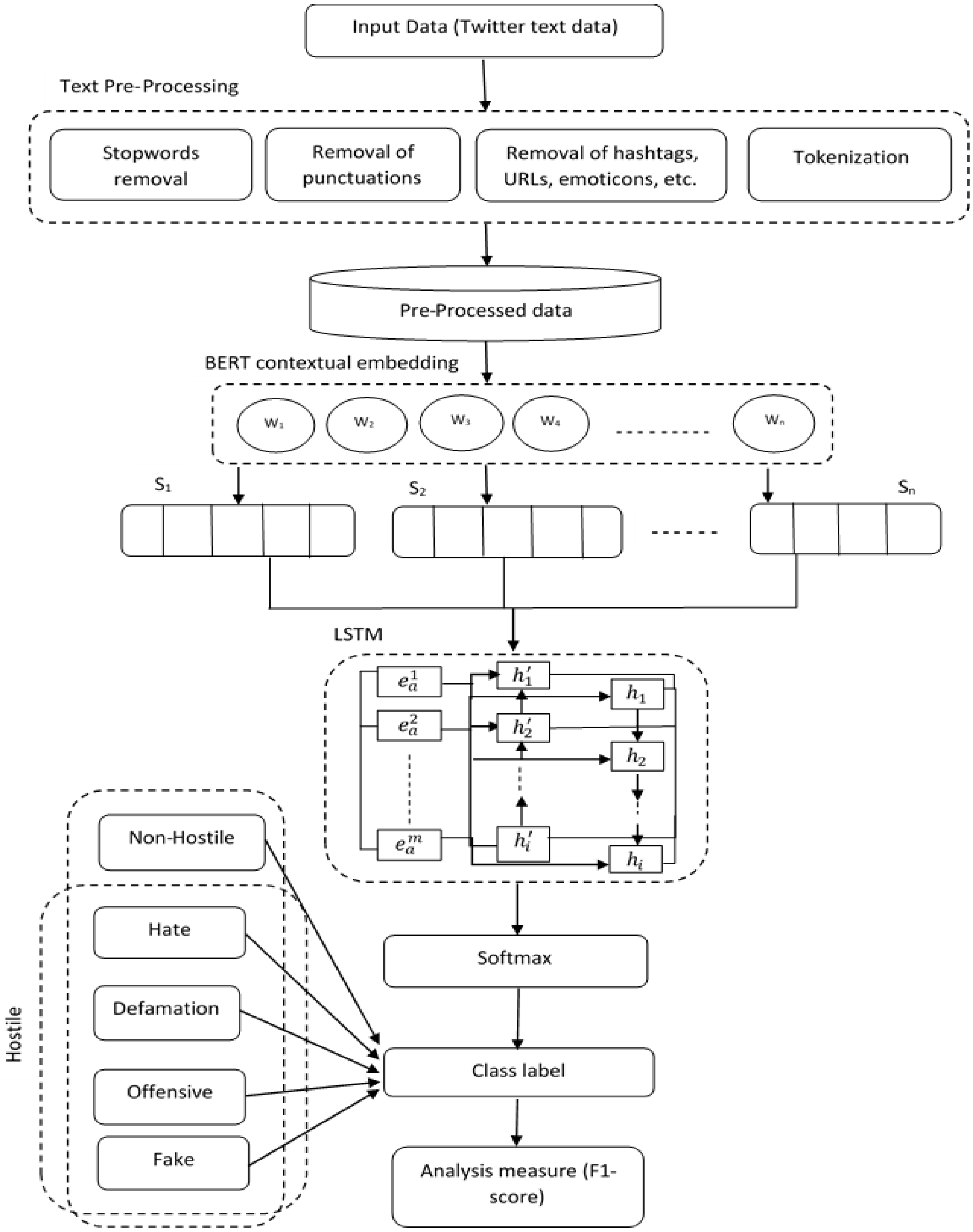


Figure 9: Detailed Architecture of Proposed approach

Table 2 states the comparison of various methods and their effectiveness metrics in classifying information into categories like non-hostile, hostile, hateful, phoney, offensive, and defamation. Linear Regression, SVM, Random Forest, MLP, BR Transformer Encoder, BR TEAP, MTDNN, and LSTM are among the models presented. The sort of language representation each model uses, such as mBert or xlm-Roberta-base, is shown in the "Embedding" column. A comparison of the proposed model's F1 score using different approaches is shown in Figure 10.

Table 2: Experimental results

Model	Embedding	Non-Hostile	Hostile			
			Hate	Fake	Offensive	Defamation
Linear Regression	mBert	83.98	44.27	68.15	38.76	36.27
SVM	mBert	84.11	47.49	66.44	41.98	43.57
Random forest	mBert	79.79	6.83	53.43	7.01	2.56
MLP	mBert	83.45	34.82	66.03	40.69	29.41
BR Transformer Encoder	xlm-roberta-base	96.12	38.62	78.96	44.84	59.31
BR TEAP	xlm-roberta-base	97.34	44.65	82.46	86.02	66.51
MTDNN	xlm-roberta-base	96.12	39.23	80.37	55.22	60.62
LSTM(Proposed work)	mBert	84.22	49.26	68.69	49.81	39.92

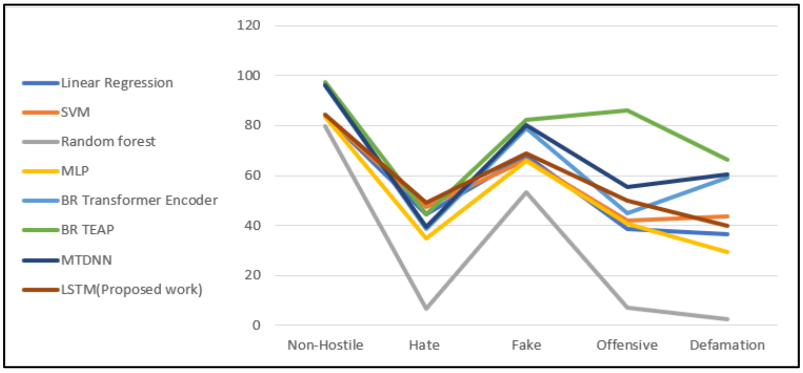
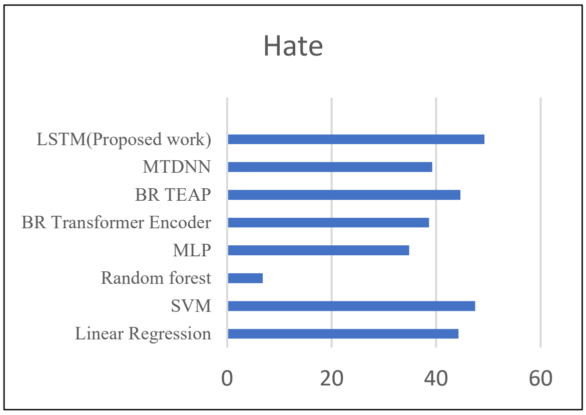


Figure 10: F1 Score Distribution of different models



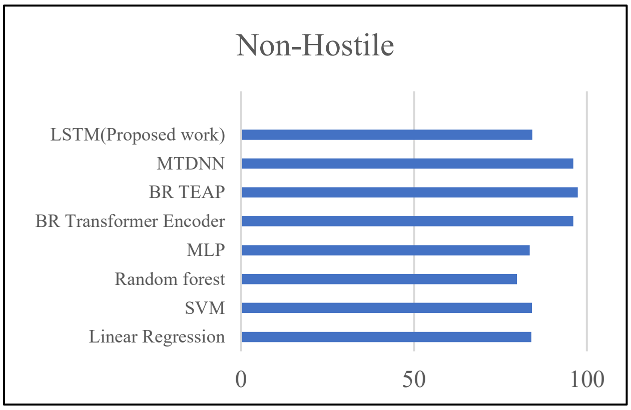


Figure 11: F1 Score Distribution of Hate class

Figure 12: F1 Score Distribution of Non- Hostile class

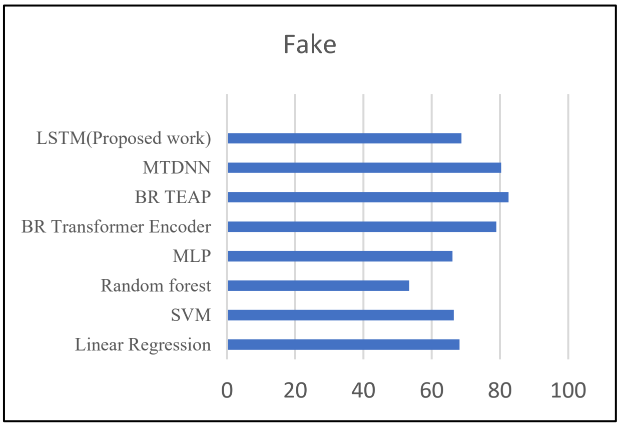


Figure 13: F1 Score Distribution of Fake class

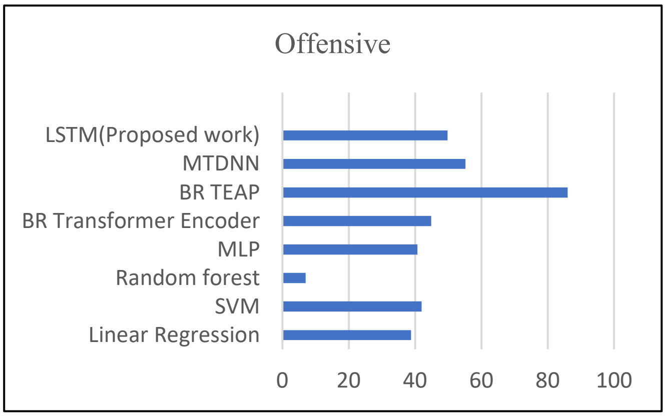


Figure 14: F1 Score Distribution of Offensive class

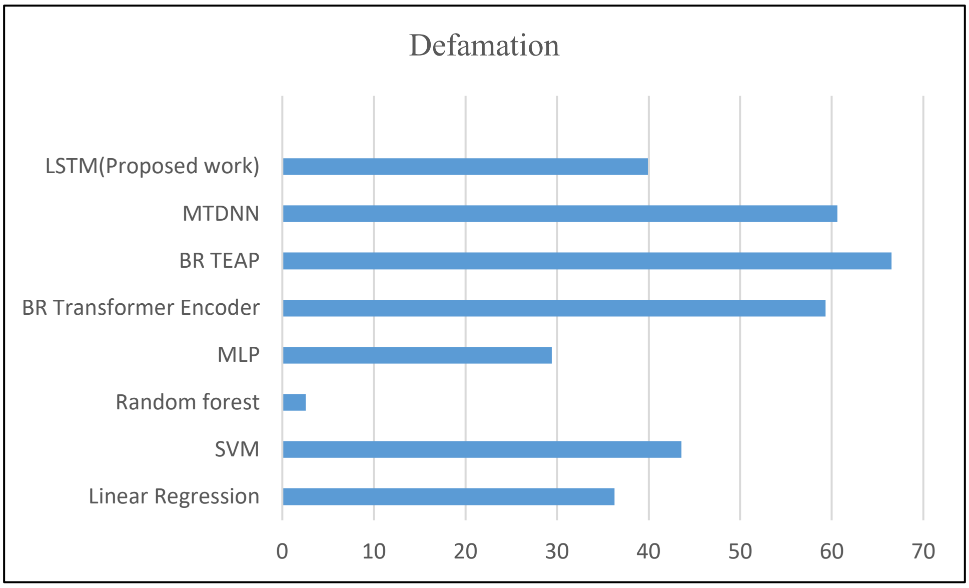


Figure 15: F1 Score Distribution of Defamation class

5.3 Discussion

This paper concludes that detecting hostile posts is a challenging problem that calls for the application of machine learning and natural language processing techniques. Many methods, including supervised learning, unsupervised learning, and deep learning, have been employed to identify hostile postings. Each strategy, though, has advantages and disadvantages. Labeled data are necessary for supervised learning algorithms to train the model, which can be time- and money-consuming [17]. Algorithms for unsupervised learning can be used to find groups of related posts, however they might not be able to identify hostile from non-hostile postings. Neural networks and other deep learning algorithms have shown excellent results in detecting aggressive posts, but they demand a lot of data and computer power.

Overall, identifying hostile messages is a difficult endeavour that necessitates a thorough comprehension of the post's linguistic and contextual characteristics. It is feasible to construct a system that is efficient at identifying hostile posts by combining various strategies and tactics, which can contribute to the creation of a safer and more encouraging online environment.

The model's training aims to minimize a loss function, but in unbalanced datasets, the majority class may dominate, prioritizing non-hostile predictions. To better understand performance, F1-score is used instead of accuracy, especially in imbalanced settings.

We understand that different models and classes have varied F1 scores. For instance, the F1 scores for the various classes are given in the "LSTM (Our approach)” as follows: 84.22 for non-hostile, 49.26 for hostile, 68.69 for hatred, 49.81 for fake, and 39.92 for offensive. Figure 11-15 illustrated the F1 score distribution of different class.

Different models and classes may have varied F1 scores for a variety of reasons, including the complexity of the models, the embeddings chosen, the characteristics of the content being classified, and the particular dataset used for training and evaluation [18].

6. Conclusion

In our study of Truculent Post Detection in Hindi, we aimed to effectively monitor and control hostile social media posts. This effort is especially significant for society as hostile word detection brings numerous benefits. It plays a vital role in ensuring online safety, reducing the prevalence of hate speech, and safeguarding vulnerable individuals by identifying and mitigating offensive and harmful content on the internet. Our approach involved understanding the use of negation words, sarcasm, and irony through LSTM models, and we harnessed the power of multilingual BERT for precise word embeddings and semantic information. Additionally, the Indic NLP library helped us accurately tokenize sentences in Hindi. Looking ahead, our work has the potential to address similar challenges in various languages, further enhancing the well-being of online communities. Our ultimate aim is to minimize the adverse effects of hostile posts on individuals, promoting a more respectful and harmonious digital environment. Striking a balance between these benefits and concerns about free speech and potential algorithm biases remains crucial and requires ongoing consideration by society and technology companies.

We perform an extensive study of the Truculent Post Detection in Hindi. The proposed model's primary goal is to monitor the social media posts and put control on the hostile ones effectively.

It plays a vital role in ensuring online safety, reducing the prevalence of hate speech, and safeguarding vulnerable individuals by identifying and mitigating offensive and harmful content on the internet. We focused on understanding the negation words, sarcasm and irony using the LSTM model, to classify the class labels accurately and build a deep-learning strategy. We benefited from multilingual BERT to generate accurate word embeddings and provide semantic information. Additionally, the Indic NLP library helped to meticulously tokenize the sentences and take in account the Hindi language. In the future, we hope this work will help to resolve comparable issues for other languages as well. Our ultimate aim is to minimize the side effects of hostile posts on the humans by analysing them and taking action accordingly.

References

[1] M. Bhardwaj, M.S. Akhtar, A. Ekbal, A. Das, T. Chakraborty, Hostility detection dataset in hindi (2020). arXiv:2011.03588.

[2] V. Bhatnagar, P. Kumar, S. Moghili, and P. Bhattacharyya, Divide and conquer: An ensemble approach for hostile post detection in hindi In Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1 (pp. 244-255). Springer International Publishing.

[3] V.K. Jha, P. Hrudya, P. Vinu, V. Vijayan, and P. Prabaharan, Dhot-repository and classification of offensive tweets in the Hindi language, Procedia Computer Science, 171 (2020) 2324–2333.

[4] S.M. Jayanthi, A. Gupta, Sj_aj@ dravidianlangtech-eacl2021: Task-adaptive pre-training of multilingual bert models for offensive language identification, arXiv preprint arXiv:2102.01051 (2021).

[5] Bhatnagar, Varad, Prince Kumar, and Pushpak Bhattacharyya. "Investigating hostile post detection in Hindi." Neurocomputing 474 (2022): 60-81.

[6] Torregrosa, Javier, Sergio D'Antonio-Maceiras, Guillermo Villar-Rodríguez, Amir Hussain, Erik Cambria, and David Camacho. "A mixed approach for aggressive political discourse analysis on Twitter." Cognitive computation 15, no. 2 (2023): 440-465.

[7] Bathla, Gourav, Pardeep Singh, Rahul Kumar Singh, Erik Cambria, and Rajeev Tiwari. "Intelligent fake reviews detection based on aspect extraction and analysis using deep learning." Neural Computing and Applications 34, no. 22 (2022): 20213-20229.

[8] Schmidt, Anna, and Michael Wiegand. "A survey on hate speech detection using natural language processing." In Proceedings of the fifth international workshop on natural language processing for social media, pp. 1-10. 2017.

[9] A.G. d'Sa, I. Illina, D. Fohr, "Bert and fasttext embeddings for automatic detection of toxic speech." In 2020 International Multi-Conference on:"Organization of Knowledge and Advanced Technologies"(OCTA), pp. 1-5. IEEE, 2020.

[10] T. Raha, S.G. Roy, U. Narayan, Z. Abid, V. Varma, "Task adaptive pretraining of transformers for hostility detection." In Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1, pp. 236-243. Springer International Publishing, 2021.

[11] R.K Singh,M.K Sachan,R.B Patel, "Cross-domain opinion classification via aspect analysis and attention sharing mechanism." Concurrency and Computation: Practice and Experience 34, no. 15 (2022): e6957.

[12] A. De, Venkatesh E, Kumar Maurya, M.S. Desarkar: "Coarse and fine-grained hostility detection in Hindi posts using fine tuned multilingual embeddings." In Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1, pp. 201-212. Springer International Publishing, 2021.

[13] Badjatiya, Pinkesh, Shashank Gupta, Manish Gupta, and Vasudeva Varma. "Deep learning for hate speech detection in tweets." In Proceedings of the 26th international conference on World Wide Web companion, pp. 759-760. 2017.

[14] Z. Waseem and D. Hovy. "Hateful symbols or hateful people? predictive features for hate speech detection on twitter." In Proceedings of the NAACL student research workshop, pp. 88-93. 2016.

[15] O. Kamal,A. Kumar ,and T. Vaidhya, "Hostility detection in Hindi leveraging pre-trained language models." In Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop,

CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1, pp. 213-223. Springer International Publishing, 2021.

[16] Hossain, M.Z., Rahman, M.A., Islam, M.S., Kar, S., "Banfakenews: A dataset for detecting fake news in bangla." arXiv preprint arXiv:2004.08789 (2020).

[17] Vinayak, S., Sharma, R., & Singh, R., "MOVBOK: A personalized social network based cross domain recommender system." Indian Journal of Science and Technology 9, no. 31 (2016): 1-10

[18] Singh, R. K., Sachan, M. K., & Patel, R. B., "Cross-domain sentiment classification using decoding-enhanced bidirectional encoder representations from transformers with disentangled attention." Concurrency and Computation: Practice and Experience 35, no. 6 (2023): 1-1.