Ensemble Classifier for Hindi Hostile Content Detection

ANGANA CHAKRABORTY, Haldia Institute of Technology, Haldia, India
SUBHANKAR JOARDAR, Haldia Institute of Technology, Haldia, India
ARIF AHMED SEKH*, School of Computer Science & Engineering, XIM UNIVERSITY, Bhubaneswar, India

Detection of hostile content from social media posts ($Facebook^{TM}$, $Twitter^{TM}$ etc.) is a demanding task in the field of Natural Language Processing (NLP). Daily growing nature of hostile content in different electronic media opened up new challenges in language understanding. It becomes more difficult in regional languages. AI-based solution is required to identify hostile content on a large scale. Though a satisfactory amount of researches has been carried out in the English language, finding hostile content in regional languages is still under progress due to unavailability of suitable datasets and tools. In terms of the number of speakers, Hindi ranks third in the world and first in the Indian Subcontinent. The objective of the article is to design hostile content detection system in Hindi language using coarse-grained (binary) classification and fine-grained (multi-class, multi-label) classification. We noted that different baseline learning method with different pre-trained language models perform differently. Using the Constraint 2021 Hindi Dataset, this research proposes a Bidirectional Encoder Representations from Transformers (BERT) based contextual embedding technique with a concatenation of emoji2vec Embedings to classify social media posts in Hindi Devanagari script as hostile or non-hostile. Additionally, for the fine-grained tasks where hostile posts are sub-categorized as defamation, fake, hate, and offensive, we develop an Ensemble Classifier varying different learning methods and embedding models. With an F1-Score of 0.9721, it is found that our proposed Indic-BERT+emoji model outperforms the baseline model and other existing models for the coarse-grained task. We have also observed that our proposed Ensemble method is giving good results than the existing models and the baseline model for the fine-grained tasks with F1-Score of 0.43, 0.82, 0.58 and 0.62 for defamation, fake, hate, and offensive classes respectively. The code and the data are available in https://github.com/skarifahmed/hostile.

CCS Concepts: • **Computing methodologies → Information extraction**.

Additional Key Words and Phrases: Hostility Detection, NLP, Social Media, Hindi, Defamation, Fake, Hate, Offensive, BERT

Authors' addresses: Angana Chakraborty, Haldia Institute of Technology, Haldia, India, angana.chakraborty9@gmail.com; Subhankar Joardar, Haldia Institute of Technology, Haldia, India, subhankarranchi@yahoo.co.in; Arif Ahmed Sekh*, School of Computer Science & Engineering, XIM UNIVERSITY, Bhubaneswar, India, skarifahmed@gmail.com.

2375-4699/2023/4-ART $15.00
https://doi.org/10.1145/3591353

1 Introduction

As of October 2021, there are 4.55 billion social media users worldwide [17]. Due to the low cost and convenient access to the internet, individuals from all over the world use social media platforms like Facebook, Twitter, Instagram, Whatsapp, etc. Due to the massive amount of individuals using social media, there is an exponential rise in the amount of user-generated material. These user-generated contents have given rise to hostile posts over social media, including spreading false rumors, trolling, offensive and hate contents, death threat, etc.

Due to the users' freedom to post, respond, and discuss on social media without restriction, a lot of hostile content is produced. In a number of instances, it has been discovered that these posts are prejudiced against a certain group of people, religion, or even a nation [44]. The number of people who support hate speech and offensive language towards Asians has increased by around 200% during the COVID-19 pandemic. The number of people who post hostile materials about Chinese people has increased by about 900%. It has been reported that there is 40% increase in toxic language used by the gaming community of and 70% spike in hate speech among teenagers and children online [9]. There have been numerous instances where hostile content has generated racial division, riots in local communities, mob lynchings, and even fatalities. Therefore, it's important to spot and stop these activities in online forums [20]. India is positioned fourth on the Social Hostilities Index [36], demonstrating the urgent need to monitor abusive online conversations. Today's greatest issue is determining what information on social media is original and what is fake.

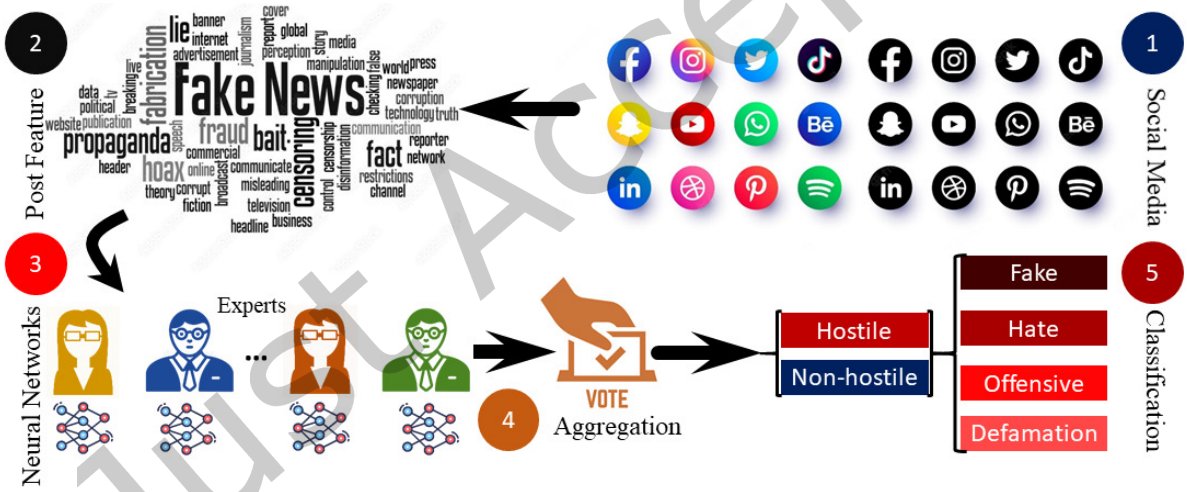

Fig. 1. The proposed pipeline for hostile content detection in social media contents. This a 5 step processed marked in circle.

1.1 Motivation

Accurate sub categorization of contents such as Fake, Hate, Offensive, Defamation, etc. is become challenging [12]. When the post is in regional languages such as Hindi, it makes the process more challenging [14]. There are 615 million active Hindi speakers worldwide, placing Hindi as the third most spoken language [29]. As numerous people uses the Devanagari (Hindi) script for interactions, so the study of the language becomes important. The unavailability of proper tools and proper datasets for said tasks in Hindi have made the language a low-resource

language [43]. So, in this article, we are motivated to choose the Hindi language for the detection of hostile posts on Social media using different neural architectures.

The detection of hostility has been carried out in two classification stages: a coarse-grained (binary) classification and a fine-grained classification (subclasses). The following is a brief explanation of them:

1. Coarse-grained classification: In this stage, each post is categorized as hostile or non-hostile. It is consequently also considered as a binary classification problem [20]. Other studies on text classification at a coarse-grained level have been conducted like Mekala et al. [32] classified news articles into five coarse-grained types using the NYT dataset (i.e. sports, arts). Additionally, they used the 20News dataset, which is a group of news articles. The documents are divided into six coarse-grained groups (i.e. computers, recreation). Zhang et al. [48] also used 20News dataset for coarse-grained classification in large volume of text information. Sentiment classification (thumbs-up and down) can also fall under coarse-grained classification [1].

2. Fine-grained classification: It is the second stage of the classification process. If the post is classified as hostile, then a more fine-grained classification has been done on hostile classes. We have performed a multi-class, multi-label classification of the hostile classes. Every hostile post is sub-categorized as defamation, fake, hate, and offensive [39]. Other research activities on text data for fine-grained classification are also discussed here. Using the NYT dataset, Mekala et al. [32] categorised news articles into 25 fine-grained classes (i.e. basketball, movies). They also used 20News dataset for fine-grained classification of news articles into 20 classes (i.e. baseball, hockey, graphics, windows). Zhang et al. [48] also used 20News dataset for fine-grained classification in large volume of text information. Das et al. [18] did fine-grained classification of insincere questions. Fine-grained classification also deals with sentiment extraction (e.g. Who offered an opinion on the sentiment, How strong is the sentiment? What is the cause of the sentiment?) [1].

1.2 Challenges

In this article, we have explored Constraint 2021 Hindi dataset. We have also proposed some methods for solving different challenges related to hostile content detection in Hindi. The primary challenges are as follows:

- Hindi is a low-resource language, hence, the lack of dataset and tools for the task is the main challenges. Despite the fact that 615 Million people speak Hindi, no effective AI technologies exist to handle Hindi language.
- Content classification in a multi-class, multi-label setup difficult in a low resource environment. Accurate identification of a large number of classes is a challenging job if the dataset is highly skewed.
- It is observed that different pre-trained model perform differently detecting different post sentiment. A particular post may be the result of various emotions. Therefore, choosing the right pre-trained model is a complicated task.

1.3 Contributions

Towards this, the contributions of the article are:

✓ Our main contributions are to incorporate state-of-the-art baselines into the Asian low-resource language, such as Hindi. We have improved the baseline by identifying the challenges and addressing the gaps. We have implemented different neural networks by varying the baseline architecture and the pre-trained language model for the task.

- ✓ We not only improve the baseline, we have also incorporated a detailed study and performance analysis by varying different embedding mechanism and feature inclusion. For the joint classification of Hindi posts into binary and multi-class, multi-label setup, we have proposed an ensemble classifier.
- ✓ We have analyzed and compared the results using state-of-the-art and recent methods, and found that the proposed methods are effective in coarse-grained and fine-grained classification.
- ✓ We have done statistical analysis of our proposed method with other recent methods.
- ✓ Some of the mistakes that our model is making on the data are analysed, and we have tried to figure out their causes. There's a scope that the performance can be increased even further by eliminating these errors.
- ✓ We have also done an Ablation Study of all our proposed methods with the baseline.

The article is organized as follows: Section 2 represents literature survey on hostile language data and neural architectures for hostile language detection in different languages. The description of the dataset used for our research is given in Section 3. Section 4 gives a description of our proposed method. Section 5 describes the experimental setup. In Section 6, we have reported our results, and we have analyzed them. Section 7 represents Ablation Study. Conclusion and Future scope are discussed in Section 8.

2 Literature Survey

Hostile Languages, Data, and Labeling: We explored the following literature revolving around hostile language detection in social media posts. Bhardwaj et al. [13] proposed a new multidimensional hostility detection dataset in the Hindi language, which was part of the CONSTRAINT-2021 shared task on hostile post detection in Hindi. The authors annotated about 8200 posts as hostile or non-hostile across Facebook and Twitter. They also assigned fine-grained hostile labels to each of the hostile posts, i.e., fake, hate, offensive, and defamation. An interesting observation is that hostile posts have a higher average number of letters per post, even if the average number of words in hostile posts is lower than in non-hostile posts. They then benchmarked the dataset using common binary classification techniques and reported the weighted F1 scores. Mollas et al. [33] and Moon et al. [34] talked about the creation of new datasets in English and Korean Language respectively to counter bias and imbalance prevailing in the majority of existing natural language datasets. They have explained in detail the process of the dataset creation and have tested them with state-of-the-art algorithms. Moon et al.[34] also showed that models trained with a bias label performed better for hate speech detection, so bias and hate are intertwined.They have used Korean Language dataset. Hossai et al. [25] proposed a dataset in Bengali language to classify different news categories into authentic and fake classes. Davidson et al. [19] demonstrated the shortcomings of lexical detection methods with low precision for hate speech tasks in English language, as they fail to distinguish between hate speech, offensive language. They created a dataset from tweets but with three labels - hate, offensive, and neither. By testing it with common algorithms, they emphasized the separation of hate speech and offensive language. Zhang et al. [49] focused on semantics for effective classification. They claimed that hateful content shows a 'long tail' pattern compared to non-hateful text due to lack of unique, discriminative linguistic features which make them difficult to classify. This causes current methods to classify most content as non-hate. They propose CNN based models to classify tweets in English language that lack discriminative features, that performed better than others. However, detecting hate content solely based on linguistic content is still a problem. Velnkar et al. [45] proposed a Marathi language dataset HASOC 2021 to classify tweets into hate and non-hate classes.Canhasi et al. [16] proposed a public dataset in Albanian language to classify news into true and fake classes using machine learning methods. Fawaid et al. [23] performed fake news detection in Bahasa Indonesia language using different deep learning models (CNN, BiLSTM, Hybrid CNN-BiLSTM, and BERT with Transformer Network) where BERT method with Transformer Network outperformed other models.

Neural Architectures for Hostile Language Detection: Badjatiya et al. [11] compared different deep learning models for multi-class detection of hate speech, where a tweet in English language can be classified as either sexist, racist, or neither. The article by Koratana et al. [30] described toxic speech and the approaches to address challenges involved in using a model at run time to detect hate speech in English language. They used the dataset published by Google Jigsaw on Kaggle, "Toxic Comment Classification Challenge". Roy et al. [40] proposed a CNN classifier that uses GloVe embedding vectors to capture semantic information. The problem statement here is to predict hate speech in a tweet in English language as soon as it is posted by the user. They proposed an end-to-end CNN system that acts as an n-gram feature extractor depending on the kernel size. Ghosh Roy et al. [41] proposed a novel approach to represent emojis and hashtags present in social media posts in English language. The paper by Abu-Farha et al. [22] presented a multitask learning framework that involves CNN, max-pooling as well as bi-LSTM layers in Arabic language. Shao et al. [42] proposed a novel multi-modal fake news detection method, named fake news detection based on multi-modal classifier ensemble. Their proposed method has the advantages of single-modal and multi-modal models both. For Indic languages (Hindi in particular), Bhatnagar et al. [15] suggested an ensemble-learning based approach for hostile speech detection in Hindi online posts. They combine techniques such as support vector machines with deep learning models to achieve competitive results. Shekhar et al. [43] used a combination of deep neural networks and XGBoost based models for coarse-grained as well as fine-grained classification of hostile posts. Kamal et al. [28] proposed a novel architecture for the hostile speech detection task in Hindi. They used the multi-label Hindi language dataset described in [13]. Various neural baseline models such as binary classifiers, multi-label classifiers, and multitask learning models are used by the authors for experimentation. To improve upon existing baseline models, the authors propose an approach called Auxiliary Task Based Binary Sub-Classification. Jha et al. [26] proposed a dataset named Devanagari Hindi Offensive Tweets (DHOT) dataset to classify tweets in Hindi language using fast-Text based model into abusive and non-abusive classes.

It is evident from the literature review that research done on hostile speech detection in the Hindi language is somewhat limited. This is surprising as Hindi is ranked among the third most widely spoken languages all over the world. Therefore, we aim to contribute to this research area by extending the work done by [28].

3 Collection of data

We have evaluated our proposed approaches using the dataset in [13] which is publicly available in CONSTRAINT 2021 shared task website [37]. This dataset consists of 8192 social media (Facebook, Twitter, etc.) posts in Hindi (Devanagari script) among which 4358 posts represent the non-hostile category and the remaining 3834 posts represent different hostile classes ('defamation', 'fake', 'hate', 'offensive', 'non-hostile').

For collection of data for each class, different approaches are followed [13].

(1) For Defamation class, different trending news stories are analyzed where some people or a group are publicly disgraced for misleading information.
(2) For collection of Fake news, some of India's leading fact-checking websites, such as BoomLive [7], Dainik Bhaskar [8] are followed. The contents of false news are determined using this technique. A topic-by-topic list of keywords is also created for each piece of false news. Later, posts from several social media environments, including Facebook, Twitter, and others are also analyzed for collection of fake posts.
(3) Tweets that incite brutality against minorities because of their race, religion, or other characteristics are examined for collection of hate speech. Different users' timelines are also reviewed where substantial amount of hate postings are uploaded. The users who commented or liked in favour of the hate speech are also analyzed, and their timelines are searched for further posts that are related to hate.
(4) To filter out offensive posts, most objectionable Hindi words which are determined by Jha et al. [26] are used. Twitter API [10] is used to extract offensive tweets containing each curse word. Each of the recorded

tweets are carefully checked for offensiveness. During the data collection process, one important finding is that offensive posts directed at women are more vulgar and hateful than their male counterparts.

(5) Posts from some reliable sites (e.g., BBCHindi) are extracted to collect non-hostile information.

A hostile post may contain multiple labels. For example a post "हमारे हिन्दू जाट भाईओ पर बोला गहलोत देख लो।। और वोट दो जाट भाईओ ये साले किसी के सगे नही है ।" can be Fake and Offensive both. Table 1 represents some annotated samples from the dataset.

Table 1. Some annotated samples from the dataset

Unique ID	Post	Labels Set
1	"अंडरवर्ल्ड डॉन छोटा राजन के भाई को बीजेपी द्वारा टिकट मिला है।"	Fake
51	"बिहार में दिखावे के लिए समय समय पर पलटू राम और फेंकू राम नूरा कुश्ती लड़ते है फिर चुनाव के बाद एक होकर भ्रष्टाचार करते है।"	Defamation
26	"हमारे हिन्दू जाट भाईओ पर बोला गहलोत देख लो।। और वोट दो जाट भाईओ ये साले किसी के सगे नही है ।"	Fake, Offensive
79	"भारत ने चीन से सीमा पर तैनात उसके सैनिकों को अनुशासन और नियंत्रण में रखने को कहा।"	Non-hostile

3.1 Details of Dataset

The annotated dataset contains 810, 1638, 1132, 1071 posts for defamation, fake, hate, and offensive classes respectively. We have split the dataset into 70%, 20%, 10% ratios for train, validation and test sets respectively. The various classes of the dataset are explained as follows [2], [13]:

- ✓ **Defamation**: Misleading information which is targeted on an individual or group to destroy their reputation publicly.
- ✓ **Fake News**: Information or claim which is verified as false.
- ✓ **Hate Speech**: Post which is targeted on a particular community of people depending on their religious beliefs, ethnicity, race, geographical belonging, etc. to spread violence or hate.
- ✓ **Offensive**: A post that is targeted on an individual or group, and it contains impolite, vulgar, profanity, or rude language to abuse an individual or group.
- ✓ **Non-Hostile**: A post that contains no hostility.

Table 2 represents the Constraint 2021 Hindi Dataset description.

Table 2. CONSTRAINT 2021 Hindi Dataset Summary

Split Percentage	Defamation	Fake	Hate	Offensive	Total Hostile Posts	Non-Hostile
Train(70%)	564	1144	792	742	2678	3050
Validation(20%)	77	160	103	110	376	435
Test(10%)	169	334	237	219	780	873
Overall	810	1638	1132	1071	3834	4358

From Table 2, we have observed that the dataset is not balanced for most hostile subclasses. For example, there are very few samples for some labels like defamation. This inherent imbalance in the dataset can result in unstable model performance. To account for this imbalance, we describe certain preprocessing and data augmentation techniques in Section 5.1.

4 Methodology

The objective of the research is to detect if a post contains hostile language or not. We have performed binary (hostile/ non-hostile), and also a multi-class, multi-label classification of hostile posts. For multi-class, multi-label

classification, hostile posts are divided into fake, hate, offensive, and defame classes. Firstly, we have considered the model approached by [28] as our baseline model. The model consists of a 4-step process as depicted in Fig. 2. Next, we use different pre-trained feature embedding mechanism to vary the representation. Finally, an ensemble classifier based on hard voting technique is proposed for multi-class, multi-label classification. Hereafter, the baseline model, embedding variation, and the proposed ensemble classifier is discussed.

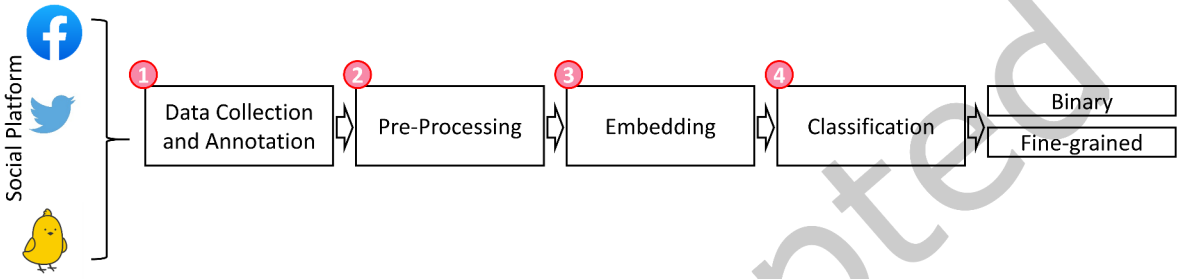

Fig. 2. Flow of the state-of-the-art classification pipeline used in the article for baseline and all other modified methods.

Baseline Model: We have chosen the model proposed by [28] as the baseline model. It takes the social media post as input text and divides it into an auxiliary/coarse-grained class (hostile/non-hostile) and different fine-grained subclasses (hostile classes). The model uses a shared BERT as pre-trained language model for end-to-end training as shown in Fig. 3. The multi-class classification is considered as "Auxiliary Task", and used in "Binary Sub-Classification". The loss is calculated based on a weighted binary cross-entropy and a multi-class log loss. The weighted log loss is defined in equation (1)[47]. The log loss is calculated for binary (L_{binary}), and multi-class ($L_{multi-class}$) both, where W_i is the weight. The final loss is defined in equation (2). $\lambda = 0.5$ when the post is hostile, otherwise set to 0. We have modified the embedding layer and also the baseline model for our experiment. We discuss the modified networks hereafter.

$$L = -W_i[y_i.\log\sigma(x_i) + (1 - y_i).\log(1 - \sigma(x_i))] \tag{1}$$

$$L_{total} = L_{binary} + \lambda.1/N.L_{multi-class} \tag{2}$$

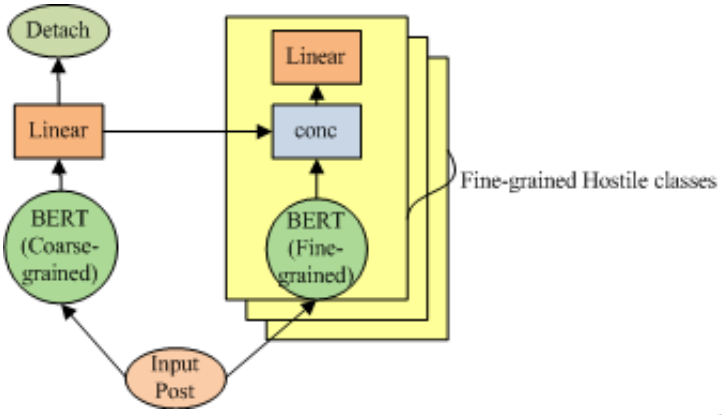

Fig. 3. Architecture of the baseline model proposed in [28]. The liner layer (fully connected) of the binary classifier is concatenated with the fully connected layer of the fine-grained classifier.

BERT with concatenated Emoji Embeddings (BERT + Emoji): The architecture of this model is similar to the coarse-grained classifier of the baseline model. Here, we have experimented with five different BERT architectures (m-BERT [3], Indic-BERT [27], HindiBERTa [35], ROBERTa Hindi [5], and Indic-Transformers Hindi XLMRoBERTa [6]). In this model, the BERT contextual embeddings are concatenated with vector representations of the emoji sequence present in the post. In the absence of emojis, a zero vector is used in place of the emoji embeddings. On the other hand, if there are multiple emojis in a post, the average vector representation is used in this case. Fig. 4 illustrates this approach.

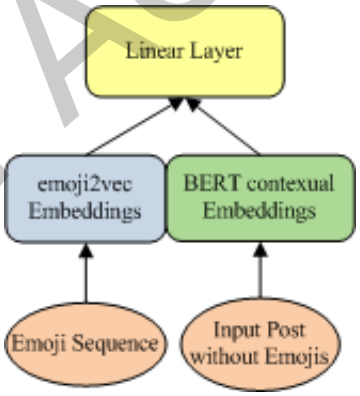

Fig. 4. The modified version of the method shown in Fig. 3. Here, BERT+Emoji is used for embeddings.

Bi-LSTM (BERT + Bi-LSTM): Here, we replace the baseline classifier with a bidirectional LSTM. This approach involves a bidirectional LSTM layer stacked on top of the BERT embeddings. The final hidden state of the sequence of tokens is fed as input to the bi-LSTM. Since the LSTM layer is bidirectional, the hidden state corresponding to the classifier token is then fed to a linear layer for classification. Fig. 5 illustrates this architecture.

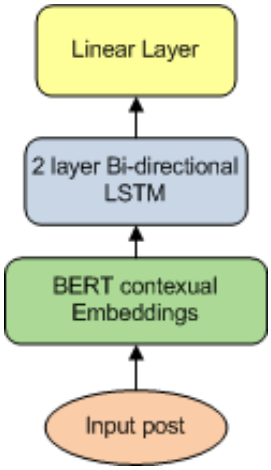

Fig. 5. Here, we have replaced the baseline by BERT+Bi-LSTM.

Ensemble Classifier: We note that different variation of the baseline model such as baseline architecture and the embedding layer affects the accuracy. Different variation performs differently for each subclasses. This motivates us to design an ensemble classifier. Using the proposed ensemble learning, we not only improve the baseline, we also incorporate a detailed study and performance analysis varying different embedding mechanism and feature inclusion. Ensemble classifier gives better performance compared to the result of individual models. In order to further improve the performance on the fine-grained tasks, an Ensemble classifier is designed using hard voting. The predicted class (y_p) is considered as the maximum voted class by all the classifiers. When we have a tie, we chose the class having maximum probability output from the classifier. We use five BERT+Emoji models and five BERT + Bi-LSTM models for our experiment. All the methods share similar pre-processing method and augmentation techniques to deal with imbalance of any hostile classes in the dataset. The pre-processing technique is discussed in Experiments section. For coarse-grained task, the proposed BERT+emoji model performs better than the baseline model. For fine-grained task, Ensemble classifier using hard voting performs better than baseline model for all the three classes (Defamation, Fake, Offensive). But, mBERT+Bi-LSTM model performs better than baseline model for hate class. We have also experimented soft voting method using average weighted technique, but soft voting method did not perform well for all the hostile classes.The algorithm for Ensemble Classifier is presented in Algorithm 1.

Algorithm 1 Ensemble Classifier

Input: $y_p, P_p : \{y_1, y_2, y_3, ..., y_N\}, P_p : \{P_1, P_2, P_3, ..., P_N\}$ ▷ *y_p is the predicted class, P_p is the predicted probability*
Output: Y_p

1: Y_p = *class having* $\max(count(y_p))$
2: **if** $(\text{count}(length(Y_p) > 1)$ **then**
3: $\quad Y_p$ = *class having* $\max(P_i)$
4: **end if**
5: Return Y_p

5 Experiments

Here, we discuss the pre-processing steps and the experimental setup in detail. The pre-processing step is applied across all the model variation.

Data pre-processing: The original dataset consists of raw $Facebook^{TM}/Twitter^{TM}$ posts and is annotated using a string of multiple classes/labels. Each post is assigned a binary label (hostile/non-hostile) and a multi-class label i.e. hostile, offensive, defame, hate, and fake.

All posts are cleaned by removing stop words, punctuation, and special characters. As we observed better performance in the presence of URLs, we decided to retain them as is. We believe that URLs add semantic value and facilitate the classification tasks. For instance, a hostile post that simply quotes content from a referenced URL, may be incorrectly classified as hostile in the absence of the URL.

Furthermore, emojis are separated from the original posts and stored as independent emoji sequences. For models requiring emoji embeddings, pre-trained emoji2vec representations were used [21]. For every emoji, a 300-dimensional vector representation was obtained using emoji2vec.

For all fine-grained tasks, the preprocessed posts are translated from Hindi to English and then each of those English language posts are back-translated into the original language (Hindi) to perform data augmentation. Data augmentation [31] is a useful technique to increase training data size and to improve the classification performance of each labels of the dataset. We have created four augmented datasets for four hostile labels (Defamation, Fake, Hate, Offensive) with increased training data. Such data augmentation strategies can be effective when the original dataset is skewed for a particular class/label.

5.1 Experimental Setup

We have used Google Colaboratory Pro tool with Tesla P100-PCIE GPU for all of our experiments. Deep learning frameworks such as PyTorch [38] and Hugging Face transformers [46] are used to develop the code-base. The evaluation metric used is macro F1 score [4]. For coarse-grained task, macro F1-score is evaluated for hostile and non-hostile classes, but for fine-grained task, it is evaluated for four hostile classes (Defamation, Fake, Hate and Offensive). The macro F1-score is evaluated as the average of class-wise F1-scores:

$$F1 = \frac{1}{N}\sum_{i=0}^{N} F1_i \tag{3}$$

Here, i represents the class index and N represents the number of classes/labels.We have experimented with different hyperparameter configurations by varying the random seed, learning rate, batch size, sequence length, number of epochs, etc.

The BERT + Emoji models were trained using a batch size of 8 and maximum sequence length of 200 for the coarse-grained task (hostile/non-hostile). Five different BERT architectures (m-BERT, Indic-BERT, HindiBERTa, ROBERTa Hindi, Indic-Transformers Hindi XLMRoBERTa) are used as pre-trained language models for this task. Using AdamW [24] as the optimizer and a learning rate of 1e-5 over 10 epochs, the fine-tuning process was performed. Other hyperparameter adjustments had a negligible impact on the model's performance.

On the other hand, two different models are implemented for the fine-grained tasks. Both architectures (BERT + Emoji and BERT + Bi-LSTM) were fine-tuned for 15 epochs, as the augmented datasets take longer to train.

Furthermore, we used a batch size of 32 and the maximum sequence length was set to 200. Optimization was performed using the AdamW optimizer with a learning rate of 1e-5.

Empirically, it was found that the Hate class performance of mBERT was superior. An ensemble classifier with the hard voting technique was implemented to further enhance results for the other fine-grained tasks (Defamation, Fake, and Offensive).

6 Results and Discussion

Here, we discuss the results obtained by different experiments and discussion of the results. We mainly reported two experiments namely binary classification task (hostile/non-hostile), and multi-class, multi-label (fine-grained). Table 3 and Table 4 represents the comparative analysis of our proposed methods with the baseline and other recent methods using Constraint 2021 Hindi Dataset for coarse grained tasks and fine-grained tasks, respectively.

Table 3. Comparative Analysis of the proposed method with other recent methods for the coarse-grained task

Model	F1-Score
Auxiliary Model [28]	0.9583
m-BERT+XGBoost [43]	0.9691
Binary Relevance Model [15]	0.9709
m-BERT+SVM [13]	0.8422
m-BERT+LR [13]	0.8398
Proposed BERT+Emoji (Indic-BERT)	**0.9721**

From Table 3, it has been seen that the Proposed Indic-BERT + Emoji model outperforms the baseline method where an auxiliary classifier has been proposed by Kamal et al. [28] and all the recent methods (m-BERT+XGBoost [43], Binary Relevance Model [15], m-BERT+SVM [13] and m-BERT+LR [13]) on the coarse-grained task. Details about baseline model is already discussed in Section 4.

Table 4. Comparative Analysis of the proposed method with other recent methods for the fine-grained tasks

Model	Defamation	Fake	Hate	Offensive
Auxiliary Model [28]	0.42	0.77	0.57	0.61
m-BERT+XGBoost [43]	0.43	0.80	0.55	0.58
Binary Relevance Model [15]	0.42	0.81	0.49	0.56
m-BERT+SVM [13]	0.39	0.68	0.49	0.41
m-BERT+LR [13]	0.36	0.68	0.44	0.38
Proposed BERT+Bi-LSTM (mBERT)	0.42	0.77	**0.58**	0.56
Proposed Ensemble Method	**0.43**	**0.82**	0.56	**0.62**

From Table 4, it is evident that for the fine-grained tasks, the proposed mBERT + Bi-LSTM model and the Proposed ensemble classifier using hard voting outperform the Auxiliary model which is considered as baseline [28]. Proposed m-BERT+Bi-LSTM is giving good result for hate class and Proposed Ensemble Method is giving good result for Defamation, Fake and Offensive classes compared to other recent methods which are listed in Table 4.

Discussion about other recent methods which are listed in Table 3 and Table 4 is given in Section 2 (Literature Survey).

6.1 Statistical Analysis of Results

In this section, a statistical analysis of the results are depicted for coarse-grained task and fine-grained tasks. We have used a box plot to demonstrate the distributions of the accuracy of different methods. It is noted that different methods having different accuracy and a large standard deviation in the terms of the F1 score. Our proposed fusion method is successfully able to put them in the top position. It signifies that the different methods having different accuracy and some of them vary excessively. Hence, relaying on a single method may not be a feasible solution for a real-world application. Our fusion enables to achieve a high accuracy in each class, and in few cases the fusion outperforms the state-of-the-art.

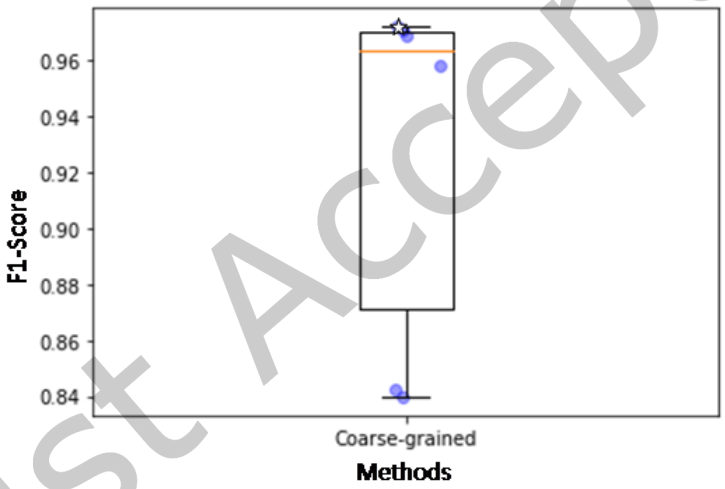

Fig. 6. Box plot for coarse-grained task. The figure shows that there exists a method who performs as minimum as 0.84, also there exist methods performing up to 0.96. It is noted that our proposed method outperforms the individual methods. Here, our proposed method result is shown with a star marker.

Fig 6 represents the box plot where coarse grained (hostile/non-hostile) results of all the methods. We observed that the standard deviation of the results are high, with a mean accuracy near to 0.96. Here, our proposed method result is shown with a star marker.

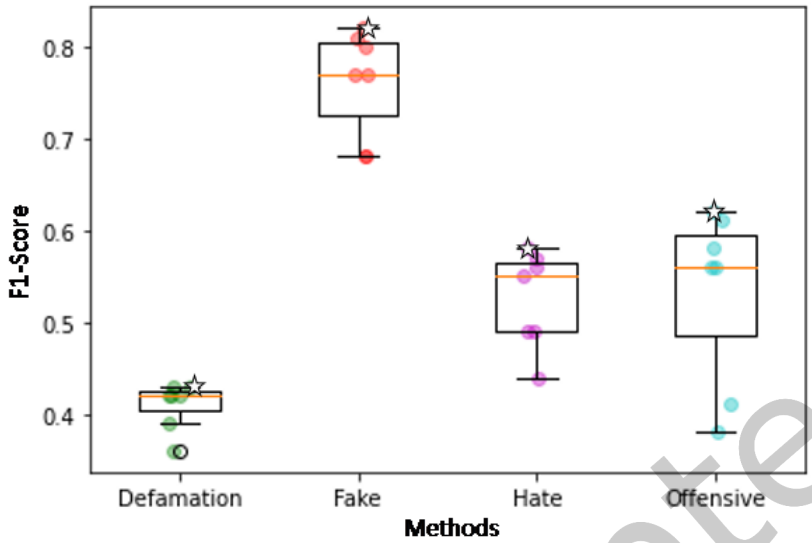

Fig. 7. Box plot for fine-grained tasks. For the defamation task, standard deviation is not so high. Our proposed method for defamation task is giving mean accuracy near to 0.42. For Fake and Hate classes our proposed ensemble method is showing mean accuracy near to 0.81 and 0.58 respectively. Standard deviation of Offensive class is high, where our proposed method is performing up to 0.61. It is noted that our proposed method outperforms the individual methods. Here, our proposed method result is shown with a star marker.

Fig 7 represents a box plot where results of all the methods including our proposed methods from Table 4 are incorporated in a box test. In fine-grained task, total four box plots are shown for four classes (Defamation, Fake, Hate, Offensive). Outputs of our proposed method are denoted by star marker for all four classes.

6.2 Training and Confusion Matrix

We observed from our experiments that different embedding methods and baseline model highly impact on loss minimization during training. Although, a particular method can perform well, still, the loss optimization may not be as good as one of the competing methods. For example, Fig. 8(b) shows one such examples (BERT+EMOJI), and Fig. 8(a) is BERT+BiLSTM, which is one of the best performing methods. It is observed that the loss optimization is much better in BERT+BiLSTM.

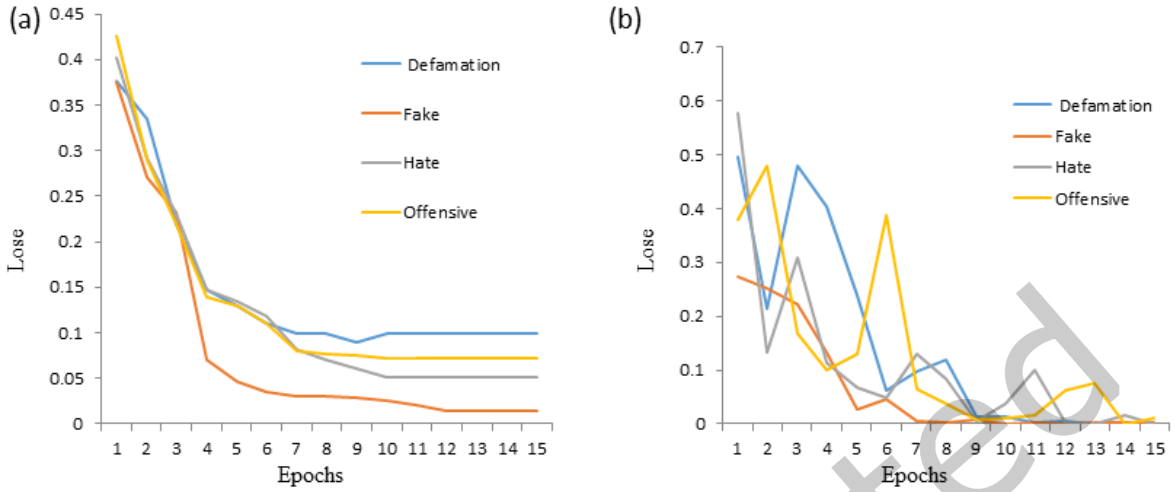

Fig. 8. Loss during training. (a) BERT+BiLSTM, (b) BERT+EMOJI.

The proposed ensemble classifier produced state-of-the-art accuracy with improvements in fine-grained classification. The proposed ensemble classifier is as good as the one of best performing methods (Fig. 9(a)) in binary classification. Fig. 9(b) shows the confusion matrix of the proposed ensemble classifier.

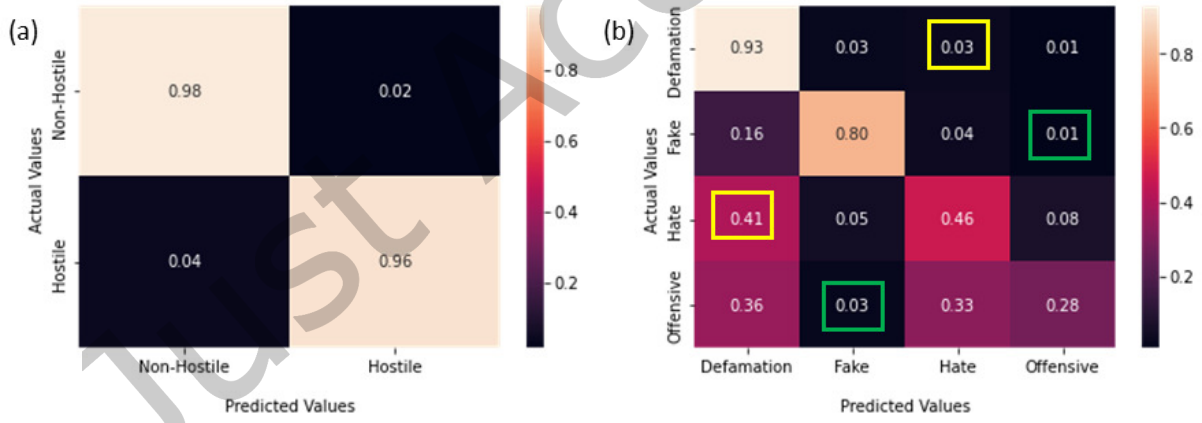

Fig. 9. The coarse-grained (binary) confusion matrix of the proposed method is proposed in (a), and the fine-grained confusion matrix of the proposed ensemble classifier is presented in (b). The best and worst performance are marked with green and yellow boxes.

6.3 Error Analysis

Our models performed well almost for maximum social media posts from the dataset, but there are some cases where our models did not perform well for some posts.Table 5 summarized true labels and predicted labels for

wrongly classified posts from the dataset using the proposed Ensemble model for different tasks (Defamation, Fake, Hate, Offensive). Appropriate justifications are also given in Table 5 for wrongly predicting some posts from the dataset.

‘

Table 5. Wrongly Classified Posts from the Dataset

Post	True Label	Predicted Label	Justification
"दोस्तों मैंने इस्तीफा देकर ही किया नां। ही किया ना? अगर हां तो रिट्वीट करें ,वरन लाइक करें।"	Fake	Not Fake	Not Fake label is detected by the model for this tweet because one can resign (इस्तीफा) from his job for different reasons.
"बंगाल के डायन के ग्रास हैं ये बेचारे, मास्टर साहब और उनका परिवार"	Defamation	Offensive	This tweet seems to be offensive because of the word "डायन".
"अंडरवर्ल्ड डॉन छोटा राजन के भाई को बीजेपी द्वारा टिकट मिला है"	Not Defamation	Defamation	This tweet seems to criticize the Government. So the model predicted as Defamation.
”@RubikaLiyaquat ये कमीनी जिस दिन मरेगी उस दिन full पार्टी“	Not Offensive	Offensive	This tweet forces the model to incorrectly predict Offensive because of the word "कमीनी".
"हमारे हिन्दू जाट भाई ओपर बोला गह लोत देख लो। और वो टदो जाट भाई ओये साले किसी के सगे नही है।"	Offensive	Hate	Hate label is detected by the model for mentioning specific word "हिन्दू" which is a religion.
"अब तक सरकार का विरोध करने वाले उछल रहे थे। उन्हें कड़ा जवाब दिया है राजपूत के लोगों ने। राजस्थानी सिर्फ दुश्मन के घर घुसकर मारते हैं बल्कि वक्त आने पर देश के साथ खड़े होना भी जानते हैं। हमें गर्व है हमारे राजस्थान और इसके वीरों पर।"	Fake	Hate	This tweet forces the model to predict Hate because it seems to use hatred word (विरोध) against Government. So understanding the real meaning is hard for the model.
”BMC की अपील: अगले सात दिनों तक गरम पानी का सेवन करें“	Fake	Not Fake	This tweet seems to be a a news or fact. So it was confusing for the model to predict as Fake.

7 Ablation Study

We conducted an ablation research for both coarse- and fine-grained tasks, comparing all of our proposed methods to the baseline method [28]. Table 6 and Table 7 highlights the proposed methods that are outperforming the baseline in terms of performance (F1-Score).

Table 6. F1-Scores for the coarse-grained task (Binary Classification)

Model	F1-Score
Baseline [28]	0.9583
Proposed BERT+Emoji (mBERT)	0.9678
Proposed BERT+Emoji (Indic-BERT)	**0.9721**
Proposed BERT+Emoji (HindiBERTa)	0.9567
Proposed BERT+Emoji (RoBERTa Hindi)	0.9583
Proposed BERT+Emoji (Indic-Transformers Hindi XLMRoBERTa)	0.9477

From Table 6, it is evident that the Proposed Indic-BERT + Emoji model outperforms the baseline model on the coarse-grained tasks. This implies that the concatenated emoji embeddings help the model for classification of a post as either hostile or non-hostile more efficiently. Since the dataset is balanced in terms of hostile/non-hostile labels (approximately 50-50 split), these results were achieved without any data augmentation techniques. Furthermore, it is worth noting that most hostile posts have emojis in them, whereas the non-hostile posts do not. Therefore, using the emoji embeddings gave a better understanding of the sentiment of the tweet and hence improved the F1 score.

Table 7. F1-Scores for the fine-grained tasks (multi-class, multi-label Classification)

Model	Defamation	Fake	Hate	Offensive
Baseline [28]	0.42	0.77	0.57	0.61
Proposed BERT+Emoji (mBERT)	0.42	0.77	0.52	0.57
Proposed BERT+Emoji (Indic-BERT)	0.39	0.78	0.51	0.60
Proposed BERT+Emoji (HindiBERTa)	0.41	0.75	0.53	0.60
Proposed BERT+Emoji (RoBERTa Hindi)	0.43	0.76	0.44	0.53
Proposed BERT+Emoji (Indic-Transformers Hindi XLMRoBERTa)	0.34	0.72	0.49	0.53
Proposed BERT+Bi-LSTM (mBERT)	0.42	0.77	**0.58**	0.56
Proposed BERT+Bi-LSTM (Indic-BERT)	0.41	0.78	0.55	0.57
Proposed BERT+Bi-LSTM (HindiBERTa)	0.41	0.76	0.50	0.60
Proposed BERT+Bi-LSTM (RoBERTa Hindi)	0.37	0.78	0.52	0.52
Proposed BERT+Bi-LSTM (Indic-Transformers Hindi XLMRoBERTa)	0.39	0.72	0.51	0.58
Proposed Ensemble Method	**0.43**	**0.82**	0.56	**0.62**

From Table 7, it is observed that for the fine-grained tasks, the proposed mBERT + Bi-LSTM model and the Proposed ensemble classifier using hard voting outperform the baseline model. In contrast, the Baseline + Emoji architecture does not produce competitive results for most fine-grained tasks. We suspect that the presence of emojis plays a part in confusing the model, as several subclasses may have identical emojis associated with them. For instance, it is hard to use emojis to differentiate between hate and offensive tweets, as both may have similar emojis. So, adding emoji embeddings ended up reducing the overall performance of our model. Stacking a Bi-LSTM layer on top of BERT did increase the performance for Hate category, as LSTM layers can be helpful in capturing additional contextual information. However, for the classification tasks for Defamation, Fake and Offensive using an ensemble gave superior results. Apart from changes in model architecture, it was observed that back-translating posts having fine-grained labels help in boosting the F1 scores.

8 Conclusion

Although our best-performing models outperform the baseline model proposed in [28], surely there is more scope for improvement. In this section, we describe potential directions for future work and ideas for improving our results. We observed our model performance stagnate due to fewer training examples and class imbalance. Augmenting the dataset by translating English tweets from the same timeline as well as region (to maintain similar cultural and geographical information) to Hindi could provide better results. Using comments on a tweet as additional data can help with coarse-grained and fine-grained tasks, both. For the coarse grained task, if comments are of similar sentiment then it should be non-hostile, whereas a hostile post may have differing and strong views. For fine-grained tasks, it is possible that comments can also mention the sentiment. For instance, someone can comment that the post is fake on a fake post. Finally, we believe that our trained models can be used to design an end-to-end pipeline for hostile language detection on social media. In the future, we will extend our work to other low-resource languages like Bengali, Marathi, etc.

References

[1] 2015. Coarse-grained vs. fine-grained sentiment analysis. Retrieved May 25, 2015 from https://www.linkedin.com/pulse/coarse-grained-vs-fine-grained-sentiment-analysis-wei-li

[2] 2021. SHARED TASK @ CONSTRAINT 2021. Retrieved Feb 08, 2021 from https://constraint-shared-task-2021.github.io/

[3] 2022. bert-base-multilingual-cased. Retrieved Jan 25, 2022 from https://huggingface.co/bert-base-multilingual-cased

[4] 2022. Documentation /Evaluation view /Classification loss metrics /Macro F1-score. https://peltarion.com/knowledge-center/documentation/evaluation-view/classification-loss-metrics/macro-f1-score

[5] 2022. flax-community/roberta-hindi. https://huggingface.co/flax-community/roberta-hindi

[6] 2022. neuralspace-reverie/indic-transformers-hi-xlmroberta. https://huggingface.co/neuralspace-reverie/indic-transformers-hi-xlmroberta

[7] 2023. BoomLive. Retrieved Jan 16, 2023 from https://hindi.boomlive.in/fake-news

[8] 2023. Dainik Bhaskar. Retrieved Jan 16, 2023 from https://www.bhaskar.com/no-fake-news/

[9] 2023. Rising Levels of Hate Speech & Online Toxicity During This Time of Crisis. Retrieved Jan 18, 2023 from https://1ight.com/Toxicity_during_coronavirus_Report-Lig.hptdf

[10] 2023. Twitter API. Retrieved Jan 16, 2023 from https://developer.twitter.com/en/docs/twitter-api

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

[12] Aditi Bagora, Kamal Shrestha, Kaushal Maurya, and Maunendra Sankar Desarkar. 2022. Hostility Detection in Online Hindi-English Code-Mixed Conversations. In *14th ACM Web Science Conference 2022*. 390–400.

[13] Mohit Bhardwaj, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Hostility detection dataset in Hindi. *arXiv preprint arXiv:2011.03588* (2020).

[14] Varad Bhatnagar, Prince Kumar, and Pushpak Bhattacharyya. 2022. Investigating Hostile Post Detection in Hindi. *Neurocomputing* 474 (2022), 60–81.

[15] Varad Bhatnagar, Prince Kumar, Sairam Moghili, and Pushpak Bhattacharyya. 2021. Divide and conquer: an ensemble approach for hostile post detection in Hindi. In *International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuation*. Springer, 244–255.

[16] Ercan Canhasi, Rexhep Shijaku, and Erblin Berisha. 2022. Albanian fake news detection. *Transactions on Asian and Low-Resource Language Information Processing* (2022).

[17] Dave Chaffey. 2022. Global social media statistics research summary 2022. Retrieved Jan 27, 2022 from https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research

[18] Sourya Dipta Das, Ayan Basak, and Soumil Mandal. 2019. Fine Grained Insincere Questions Classification using Ensembles of Bidirectional LSTM-GRU Model.. In *FIRE (Working Notes)*. 473–481.

[19] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11. 512–515.

[20] Arkadipta De, Venkatesh Elangovan, Kaushal Kumar Maurya, and Maunendra Sankar Desarkar. 2021. Coarse and fine-grained hostility detection in Hindi posts using fine tuned multilingual embeddings. In *International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuation*. Springer, 201–212.

[21] Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. 2016. emoji2vec: Learning emoji representations from their description. *arXiv preprint arXiv:1609.08359* (2016).
[22] Ibrahim Abu Farha and Walid Magdy. 2020. Multitask learning for Arabic offensive language and hate-speech detection. In *Proceedings of the 4th workshop on open-source Arabic corpora and processing tools, with a shared task on offensive language detection*. 86–90.
[23] Jibran Fawaid, Aisyah Awalina, Rifky Yunus Krisnabayu, and Novanto Yudistira. 2021. Indonesia's Fake News Detection using Transformer Network. In *6th International Conference on Sustainable Information Engineering and Technology 2021*. 247–251.
[24] Fabio M. Graetz. 2018. Why AdamW matters. Retrieved June 03, 2018 from https://towardsdatascience.com/why-adamw-matters-736223f31b5d
[25] Md Zobaer Hossain, Md Ashraful Rahman, Md Saiful Islam, and Sudipta Kar. 2020. Banfakenews: A dataset for detecting fake news in bangla. *arXiv preprint arXiv:2004.08789* (2020).
[26] Vikas Kumar Jha, Pa Hrudya, PN Vinu, Vishnu Vijayan, and Pa Prabaharan. 2020. DHOT-repository and classification of offensive tweets in the Hindi language. *Procedia Computer Science* 171 (2020), 2324–2333.
[27] Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 4948–4961.
[28] Ojasv Kamal, Adarsh Kumar, and Tejas Vaidhya. 2021. Hostility detection in hindi leveraging pre-trained language models. In *International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuation*. Springer, 213–223.
[29] Anna Klappenbach. 2022. The 12 most spoken languages in the world. Retrieved Jan 07, 2022 from https://blog.busuu.com/most-spoken-languages-in-the-world/
[30] Animesh Koratana and Kevin Hu. 2018. Toxic Speech Detection. In *Proceedings of the 32nd international conference on Neural Information Processing Systems*. 1–9.
[31] Edward Ma. 2019. Data Augmentation in NLP. Retrieved Apr 12, 2019 from https://towardsdatascience.com/data-augmentation-in-nlp-2801a34dfc28
[32] Dheeraj Mekala, Varun Gangal, and Jingbo Shang. 2021. Coarse2Fine: Fine-grained text classification on coarsely-grained annotated data. *arXiv preprint arXiv:2109.10856* (2021).
[33] Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2020. Ethos: an online hate speech detection dataset. *arXiv preprint arXiv:2006.08328* (2020).
[34] Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. BEEP! Korean corpus of online news comments for toxic speech detection. *arXiv preprint arXiv:2005.12503* (2020).
[35] M.Romero. 2022. mrm8488/HindiBERTa. https://huggingface.co/mrm8488/HindiBERTa
[36] Samir Nazareth. 2019. Removed from reality. Retrieved Feb 15, 2019 from https://www.thehindu.com/opinion/op-ed/removed-from-reality/article26272904.ece?homepage=true
[37] parthpatwa. 2022. Constraint@AAAI2021 - Hostile Post Detection in Hindi. Retrieved Feb 22, 2022 from https://competitions.codalab.org/competitions/26654#learn_the_details-dataset
[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).
[39] Parth Patwa, Mohit Bhardwaj, Vineeth Guptha, Gitanjali Kumari, Shivam Sharma, Srinivas Pykl, Amitava Das, Asif Ekbal, Md Shad Akhtar, and Tanmoy Chakraborty. 2021. Overview of constraint 2021 shared tasks: Detecting english covid-19 fake news and hindi hostile posts. In *International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuation*. Springer, 42–53.
[40] Pradeep Kumar Roy, Asis Kumar Tripathy, Tapan Kumar Das, and Xiao-Zhi Gao. 2020. A framework for hate speech detection using deep convolutional neural network. *IEEE Access* 8 (2020), 204951–204962.
[41] Sayar Ghosh Roy, Ujwal Narayan, Tathagata Raha, Zubair Abid, and Vasudeva Varma. 2021. Leveraging multilingual transformers for hate speech detection. *arXiv preprint arXiv:2101.03207* (2021).
[42] Yi Shao, Jiande Sun, Tianlin Zhang, Ye Jiang, Jianhua Ma, and Jing Li. 2022. Fake News Detection Based on Multi-Modal Classifier Ensemble. In *Proceedings of the 1st International Workshop on Multimedia AI against Disinformation*. 78–86.
[43] Chander Shekhar, Bhavya Bagla, Kaushal Kumar Maurya, and Maunendra Sankar Desarkar. 2021. Walk in Wild: An Ensemble Approach for Hostility Detection in Hindi Posts. *arXiv preprint arXiv:2101.06004* (2021).
[44] Shishir Tiwari and Gitanjali Ghosh. 2014. Social media and freedom of speech and expression: Challenges before the Indian law. *Available at SSRN 2892537* (2014).
[45] Abhishek Velankar, Hrushikesh Patil, Amol Gore, Shubham Salunke, and Raviraj Joshi. 2021. Hate and Offensive Speech Detection in Hindi and Marathi. *arXiv preprint arXiv:2110.12200* (2021).
[46] T Wolf, L Debut, V Sanh, J Chaumond, C Delangue, A Moi, P Cistac, T Rault, R Louf, M Funtowicz, et al. 2020. HuggingFace's Transformers: State-of-the-art natural language processing. arXiv 2019. *arXiv preprint arXiv:1910.03771* (2020).

[47] Tong Zeng and Daniel E Acuna. 2020. Modeling citation worthiness by using attention-based bidirectional long short-term memory networks and interpretable models. *Scientometrics* 124, 1 (2020), 399–428.
[48] Weifan Zhang, Hui Zhang, Yuan Zuo, and Deqing Wang. 2015. Modeling Both Coarse-Grained and Fine-Grained Topics in Massive Text Data. In *2015 IEEE First International Conference on Big Data Computing Service and Applications*. IEEE, 378–383.
[49] Ziqi Zhang and Lei Luo. 2019. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web* 10, 5 (2019), 925–945.