Hate speech detection: a comparison of mono and multilingual transformer model with cross-language evaluation

Koyel Ghosh
Central Institute of Technology,
Kokrajhar, Assam, India
ghosh.koyel8@gmail.com

Apurbalal Senapati
Central Institute of Technology,
Kokrajhar, Assam, India
a.senapati@cit.ac.in

Abstract

Warning: This paper contains examples of the language that some people may find offensive.

Transformer-based Language models have achieved state-of-the-art performance on a wide range of Natural Language Processing (NLP) tasks. This work will examine the effectiveness of transformer language models like BERT, RoBERTa, ALBERT, and DistilBERT on existing Indian hate speech datasets such as HASOC-Hindi (2019), HASOC-Marathi (2021) and Bengali Hate Speech (BenHateSpeech) over binary classification. Most deep learning methods fail to recognize a hate sentence if hate words are wrapped into sophisticated words where transformers understand the context of a hate word present in a sentence. Here, Transformer-based multilingual models such as MuRILBERT, XLM-RoBERTa, etc. are compared with monolingual models like NeuralSpace-BERTHi (Hindi), MahaBERT (Marathi), BanglaBERT (Bengali), etc. It is noticed that the monolingual MahaBERT model performs the best on HASOC-Marathi, whereas the multilingual MuRILBERT performs the best on HASOC-Hindi and BenHateSpeech. Several other cross-language evaluations over Marathi and Hindi monolingual models and mixed observations are presented.

1 Introduction

According to the Cambridge Dictionary, hate speech is defined as - “Hate speech is a public speech that expresses hate or encourages violence towards a person or group based on race, religion, sex, or sexual orientation”[1]. Statistics reveal that half of the world’s population, including print media, is now engaged in social media platforms[2] and 12½ trillion hours spent online by the users[3]. This trend shall continue till obvious infinity. Sometimes aggressive posts, misleading news, and harassing comments can lead people to social violence, even riots (Laub, 2019). Worldwide, Governments are introducing laws against hate speech. So, digital media like Twitter, Facebook, etc., are also becoming more concerned about it and endeavouring to filter hate, sexual abuse, harmful acts, harassment, bullying, child abuse, etc.

Researchers explore this field, but most of the experiment is based on European language datasets[4]. Limited work is done on the Indian languages except for publishing datasets or improving accuracy. India has 22 official languages and about 1,000 living languages from various language groups (Kalra and Dutt, 2019). People in India use their native lan-

[1] https://dictionary.cambridge.org/dictionary/english/hate-speech
[2] https://datareportal.com/reports/digital-2021-global-overview-report
[3] https://datareportal.com/reports/digital-2022-global-overview-report
[4] https://hatespeechdata.com/

guages on social media platforms, and sometimes users don't follow the proper structure or grammar, making it more complicated to detect hate speech in the computational aspect. This situation motivated us to work on hate speech on Twitter and other social media texts. It is challenging for automatic approaches to detect hate speech in text.

Researchers use state-of-the-art transformer models more in language-related researchs like NLP, Information Retrieval (IR), etc., to enrich performance. Many works have already been done in NLP like text classification (Sun et al., 2019), question-answering (McCarley et al., 2019), token classification (Ulčar and Robnik-Šikonja, 2020), and Named Entity Recognition (NER) (Luoma and Pyysalo, 2020). The pre-trained BERT-based masked language models have been used, and these language models' multilingual and monolingual variants have drawn attention to the low-resource languages.

This paper attempted to identify hate speech content in Hindi, Marathi and Bengali comments collected from social media. We choose various publicly available datasets like HASOC (Hate Speech and Offensive Content Identification)[5] and Bangla Hate Speech datasets (BenHateSpeech)[6] with binary classification. Variation of BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), RoBERTa (Robustly optimized BERT) (Liu et al., 2019), ALBERT (A Lite BERT) (Lan et al., 2019), DistilBERT (Distilled version of BERT) (Sanh et al., 2019) and their pre-trained models such as mBERT (Devlin et al., 2018), MuRIL-BERT (Khanuja et al., 2021), NeuralSpace-BERTHi (Jain et al., 2020), RoBERTa-Hindi, Indic Bert (Kakwani et al., 2020), MahaBERT (Joshi, 2022a), MahaRoBERTa (Joshi, 2022b), XLM-RoBERTa (Conneau et al., 2019), BanglaBert (Sarker, 2020) etc. is used for this work.

Later, we compare different pre-trained BERT architecture's performance on publicly available datasets in Hindi, Marathi and Bengali languages. We also compare multilingual and monolingual variants of these language models. The monolingual models used here are only pre-trained on Hindi, Bengali, or Marathi data. Next, we follow a cross-language evaluation of these BERT models (Litake et al., 2022) since both Hindi and Marathi share the Devanagri script.

Our focus in this work is :

- To see how well pre-trained BERT models perform, utilize the mono and multilingual pre-trained BERT models with their variants.
- A detailed comparison between all the models for Hindi, Marathi and Bengali languages has been made. Almost thirty experiments have been done, twelve for Hindi, Marathi and six for Bengali.
- There is a cross-language experiment between Hindi and Marathi where monolingual Marathi models, i.e. MahaBert, MahaRoBERTa, RoBERTa-Base-Mr, and MahaAlBERT performs well on the Hindi dataset. NeuralspaceBERTHi, Roberta-Hindi and DistilBERTHindi, which are monolingual Hindi models, also perform well in the case of the Marathi dataset.

The rest of the paper is structured as follows. Section 2 is the work related to hate speech detection in Indian languages. Section 3 describes the experimental setup like the dataset, preprocessing steps, BERT variants and pre-trained models. Section 4 summarizes the results and findings from all of the experiments. Finally, it is concluded in Section 5.

[5] https://hasocfire.github.io/hasoc/2022/index.htm

[6] https://www.kaggle.com/naurosromim/bengali-hate-speech-dataset

2 Related Work

There is very little research on the Indian hate speech dataset as less data is available publicly. Creating labelled datasets of hate speech in the Indian language is tedious and challenging. It needs lots of groundwork and preprocessing, like cleaning, annotators' agreements, etc., to create valuable data from social media. In this section, language-wise, we shall discuss some existing datasets and the work done on those datasets.

Hindi: HASOC (Hate Speech and Offensive Content Identification), a shared task organized by FIRE (Forum for Information Retrieval Evaluation)[7], which published hate datasets in Indian languages such as Hindi, Marathi, etc. HASOC offers four subtracks, one of which is relevant to us: **HASOC - English and Indo-Aryan Languages**. Datasets are distributed in tab-separated format. HASOC and most other collections require mechanisms to detect hateful content from the text of a post.

In 2019, the HASOC-Hindi dataset offered three tasks (Mandl et al., 2019). The first task is binary classification, i.e. subtask A. The second task is to find whether the hate comment was profane or abusive (multiclass), i.e. subtask B. The third is to predict whether the hate comment is targeted or untargeted (multiclass), i.e. subtask C. In the Hindi language, ninety-three runs were submitted across three subtasks. Regarding the Hindi subtask A, the winner team, QutNocturnal (Bashar and Nayak, 2020), employed a CNN base technique with Word2vec embedding and got better Marco F1 and Weighted F1 values, 0.8149 and 0.8202, respectively. The second team LGI2P (Mensonides et al., 2019), trained a fastText model for the proposed Hindi language and later used BERT for classification. The system achieved 0.8111 Marco-F1 and 0.8116 Weighted-F1 values. For sub-task B on Hindi Dataset, 3Idiots (Mishra and Mishra, 2019) scores 0.5812 and 0.7147 in Marco-F1 and Weighted-F1 utilizing BERT. Team A3-108 (Mujadia et al., 2019) achieves a high Marco-F1 score on subtask C Hindi Dataset, which is 0.5754. According to them, Adaboost (Freund and Schapire, 1997) was the best performing classifier among the three classifiers, i.e., Adaboost or Adaptive Boosting (AB), Random Forest (RF), Linear Support Vector Machine (SVM). They merge multiple weak classifiers to construct a robust prediction model, but an ensemble of SVM, Random Forest, and Adaboost with hard voting performed even better. This classifier used TF-IDF features of word unigrams and characters 2, 3, 4, and 5 grams with an additional feature of the length of every tweet.

In HASOC 2020, two Hate Speech detection tasks (Mandl et al., 2020), sub-task A (binary class) and sub-task B (multiclass), are proposed with another Hindi dataset in the research area. NSIT_ML_Geeks (Raj et al., 2020) outperforms other teams in the competition scoring Marco-F1 0.5337 and 0.2667 in sub-task A and sub-task B, respectively utilizing CNN and BiLSTM. Nohate (Kumari) team achieved Marco-F1 0.3345 in sub-task B, fine-tuning BERT model for the classification.

In 2021, HASOC published a Hindi dataset (Modha et al., 2021) with sub-task A and B again. Total Sixty-five teams submitted a total of six thousand and fifty-two runs. The best submission was achieved Macro F1 0.7825 in sub-task A with a fine-tuned Multilingual-BERT (20 epochs) with a classifier layer added at the final phase. The second team also fine-tuned Multilingual-BERT and scored Macro F1 0.7797. NeuralSpace (Bhatia et al., 2021) got Macro F1 0.5603 in sub-task B. They use an XLM-R transformer, vector representations for emojis using the system Emoji2Vec, and sentence embeddings for hash-

[7] http://fire.irsi.res.in/fire/2022/home

tags. After that, three resulting representations were concatenated before classification.

In the paper (Bhardwaj et al., 2020) they used the pre-trained multilingual BERT (m-BERT) model for computing the input embedding on the Hostility Detection Dataset (Hindi) later SVM, Random-Forest, Multilayer perceptron (MLP), Logistic Regression models are used as classifiers. In coarse-grained evaluation, SVM reported the best weighted-F1 score of 84%, whereas they obtained 84%, 83%, and 80% weighted-F1 scores for LR, MLP, and RF. In fine-grained evaluation, SVM has the most excellent F1 score for evaluating three hostile dimensions, namely Hate (47%), Offensive (42%), and Defamation (43%). Logistic Regression beats the others in the Fake dimension with an F1 score of 68%.

Marathi: In HASOC-Marathi (Modha et al., 2021), the best-performing team WLV-RIT fine-tuned XLM-R Large model with a simple softmax layer. Later executed transfer learning from English data released for OffensEval 2019 (Zampieri et al., 2019) and Hindi data released for HASOC 2019 (Mandl et al., 2019) and show that executing transfer learning from Hindi is better than executing transfer learning from English. They Scored an F1 score of 0.9144 (Nene et al., 2021). The second team applied a fine-tuned LaBSE transformer (Feng et al., 2020) on the Marathi data set and the Hindi data set and achieved an F1 score of 0.8808. Their experiments show that the LaBSE transformer (Glazkova et al., 2021) outperforms XLM-R in the monolingual settings, but XLM-R performs better when Hindi and Marathi data are merged. L3CubeMahaHate (Velankar et al., 2022) presents the first major Marathi hate speech dataset with 25,000 distinct tweets from Twitter, later annotated manually, and labelled them into four major classes, i.e. hate, offensive, profane, and not. Finally, they use CNN, LSTM, and Transformers. Next, they explore monolingual and multilingual variants of BERT like MahaBERT, IndicBERT, mBERT, and xlm-RoBERTa and show that monolingual models perform better than their multilingual counterparts. Their MahaBERT (Joshi, 2022a) model provides the best results on L3Cube-MahaHate Corpus.

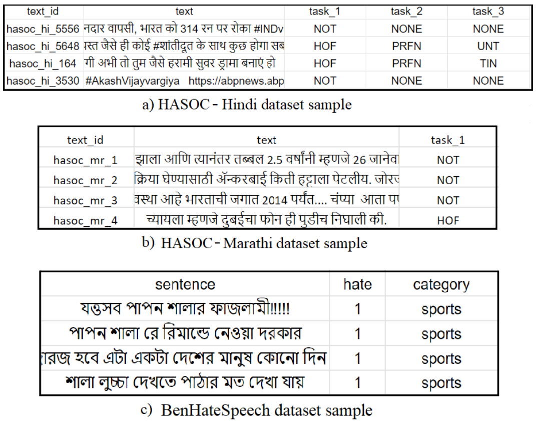
Bengali: Karim et al. (Karim et al., 2020) published a Bengali dataset with 35,000 hate statements (political, personal, geopolitical, and religious) and applied a multichannel CNN and LSTM-based approach. Later DeepHate-Explainer (Karim et al., 2021) added more than 5,000 labelled examples with it and used an ensemble method of transformer-based neural architectures to classify them into political, personal, geopolitical, and religious hates and achieved F1-scores of 78%, 91%, 89%, and 84%, for political, personal, geopolitical, and religious hates. In the paper (Romim et al., 2021), They published a Bengali Hate Speech corpus with 30,000 comments labelled with “1” for hate comments; otherwise, “0”. This paper (Mandal et al., 2022) created a political news corpus and then developed a keyword or phrase-based hate-speech identifier using a semi-automated approach.

Most of the top results are delivered by the systems based on Deep neural models and transformers.

3 Experimental Setup

3.1 Dataset Selection

The experiment uses the HASOC-Hindi (2019), HASOC-Marathi (2021) and BenHate-Speech (Romim et al., 2021) datasets. Statistics and class distribution for the training set of HASOC-Hindi, HASOC-Marathi and Ben-HateSpeech datasets are in table 1 and for the test set in table 2. Figure 1 shows samples of datasets, and we chose only binary classification task for our work, i.e., to detect whether a sentence or text conveys hate or not. For HASOC-Hindi and HASOC-Marathi, classes are “HOF” and “NOT” whereas in BenHateSpeech, classes are “1” (hate) and “0” (not).



text_id	text	task_1	task_2	task_3
hasoc_hi_5556	नदार वापसी, भारत को 314 रन पर रोका #INDv	NOT	NONE	NONE
hasoc_hi_5648	स्त जैसे ही कोई #शांतीदूत के साथ कुछ होगा सब	HOF	PRFN	UNT
hasoc_hi_164	गी अभी तो तुम जैसे हरामी सुवर ड्रामा बनाएं हो	HOF	PRFN	TIN
hasoc_hi_3530	#AkashVijayvargiya https://abpnews.abp	NOT	NONE	NONE

a) HASOC - Hindi dataset sample

text_id	text	task_1
hasoc_mr_1	झाला आणि त्यानंतर तब्बल 2.5 वर्षांनी म्हणजे 26 जानेवा	NOT
hasoc_mr_2	क्रिया घेण्यासाठी अँन्करबाई किती हट्टाला पेटलीय. जोरज	NOT
hasoc_mr_3	वस्था आहे भारताची जगात 2014 पर्यंत.... चंप्या आता प	NOT
hasoc_mr_4	च्यायला म्हणजे दुबईचा फोन ही पुडीच निघाली की.	HOF

b) HASOC - Marathi dataset sample

sentence	hate	category
যত্তসব পাপন শালার ফাজলামী!!!!!	1	sports
পাপন শালা রে রিমান্ডে নেওয়া দরকার	1	sports
ারজ হবে এটা একটা দেশের মানুষ কোনো দিন	1	sports
শালা লুচ্চা দেখতে পাঠার মত দেখা যায়	1	sports

c) BenHateSpeech dataset sample

Figure 1: Samples of HASOC-Hindi, HASOC-Marathi and BenHateSpeech datasets respectively

3.2 Preprocessing

Before employing data to the transformers, text data need to be cleaned and noise-free to enrich the performance. India’s low-resourced languages share several characteristics. Despite being written in different languages, researchers employed nearly identical preprocessing approaches for all datasets. Text preprocessing procedures can be slightly different depending on the task and the dataset used. Some datasets had raw comments with emojis, punctuation, and unwanted characters. In most cases, the following steps are used:

Normalization: It defines the removal of existing emojis, unwanted characters and stopwords from the sentences.

Removing punctuation and number: Punctuation and numbers often don’t add extra meaning to the text hence being removed from the text.

Word tokenization : Here converting a sentence into an individual word is called a token.

Stemming: Stemming removes the inflections from each word to convert that word to its root word.

Label encoding: In HASOC-Hindi and HASOC-Marathi, task_1 is tagged as “NOT” and “HOF”. We encode them into a unique number. Like, “NOT” to “0” and “HOF” to “1”, where we leave BenHateSpeech dataset as it is.

In the case of HASOC-Marathi and BenHateSpeech datasets, we followed above mentioned simple steps. In case of HASOC-hindi, we followed preprocessing techniques as mentioned in paper (Bashar and Nayak, 2020) like replacing person occurrence (e.g. @someone) with `xxatp`, URL occurrence with `xxurl`, source of modified retweet with `xxrtm`, source of not modified retweet with `xxrtu`, fixing the repeating characters (e.g. goooood), removed common invalid characters (e.g. $< br =>$, $< unk >$, $@-@$,etc) and a lightweight stemmer for Hindi language (Ramanathan and Rao, 2003) for stemming the words.

3.3 Transformer Language Models

Figure 2 shows a general transformer-based BERT model structure, and the input text is in the Bengali language. After the above steps of preprocessing, we employ the p number of texts of the training set (D) is indicated as

$$D = \{T_1, T_2, T_3, .., T_i, ...T_p\},$$

where T_i is the i^{th} number of texts and p is equal to the total number of texts present in a training set. Given a text T_i, the text having m words, i.e., length of the text, is denoted as

$$T_i = \{w_{i,1}, w_{i,2}, w_{i,3}, ..., w_{i,k}, .., w_{i,m}\},$$

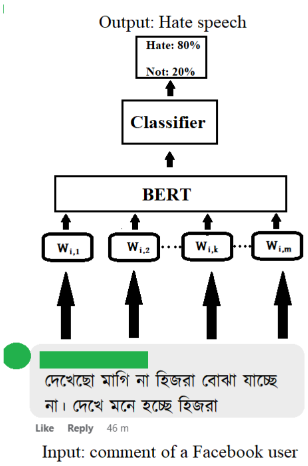
where $w_{i,k}$ denotes the k^{th} word in the i^{th} text.

Datasets	HOF/Hate	NOT	Total
HASOC-Hindi (2019)	2,469	2,196	4,665
HASOC-Marathi (2021)	669	1,205	1,874
BenHateSpeech	8,000	16,000	24,000

Table 1: Class distribution analysis for training set

Datasets	HOF/Hate	NOT	Total
HASOC-Hindi (2019)	605	713	1318
HASOC-Marathi (2021)	207	418	625
BenHateSpeech	2,000	4,000	6,000

Table 2: Class distribution analysis for test set



দেখেছো মাগি না হিজরা বোঝা যাচ্ছে না। দেখে মনে হচ্ছে হিজরা

Figure 2: A general transformer-based BERT model architecture

3.3.1 BERT

BERT is developed by Google, a transformer-based technique for NLP. BERT can generate contextualized embeddings. It produces almost similar vectors for synonyms and different vectors if the use of words is different. During training, it learns the details from both sides of the word's context. So, it is called a bidirectional model. We evaluated mono and multilingual BERT on Hindi, Marathi and Bengali datasets. Due to memory and *GPU* issues, we did several experiments but with the same hyperparameter combination (Table 3).

mBERT[8]: It is pre-trained with the largest Wikipedia over 104 top languages worldwide, including Hindi, Bengali and Marathi, using a masked language modelling (MLM) objective.

MuRILBERT[9]: Multilingual Representations for Indian Languages (MuRIL) is a BERT model pre-trained on 17 Indian languages and their transliterated counterparts, i.e. monolingual segments and parallel segments.

NeuralspaceBERTHi[10]: This BERT model is pre-trained on approx. 3 GB of monolingual training corpus, i.e., OSCAR corpus released by neuralspace-reverie. It fine-tuned downstream tasks like text classification, POS-tagging, question-answering, etc.

MahaBERT[11]: MahaBERT is a multilingual BERT (bert-base-multilingual-cased) model finetuned on L3Cube-MahaCorpus and other publicly available Marathi monolingual datasets.

BanglaBERT[12]: Using mask language modelling, bangla-Bert-Base was pre-trained on data downloaded from OSCAR and Bengali Wikipedia Dump Dataset.

[8] https://huggingface.co/bert-base-multilingual-cased

[9] https://huggingface.co/google/muril-base-cased

[10] https://huggingface.co/neuralspace-reverie/indic-transformers-hi-bert

[11] https://huggingface.co/l3cube-pune/marathi-bert

[12] https://huggingface.co/sagorsarker/bangla-bert-base

Hyperparameter	BERT variants
Learning-rate	$1e\text{-}5$
Epochs	5
Max seq length	512
Batch size	8

Table 3: Combination of hyperparameters for training BERT variants

3.3.2 RoBERTa

More extended time training on a large dataset can increase BERT's performance. This model, called RoBERTa, a self-supervised transformer model trained on raw texts, outperforms BERT by 4%-5% on natural language inference and utilizes a character-level BPE (Byte Pair Encoding) tokenizer. Still, RoBERTa uses a byte-level BPE tokenizer, which benefits from a universal encoding scheme.

XLM-RoBERTa (base-sized model)[13]**:** The XLM-RoBERTa model is a multilingual version of the RoBERTa model pre-trained on 2.5 TB of filtered CommonCrawl data containing 100 languages. It does not require $lang$ tensors like XLM multilingual models to determine which language is utilized and choose the correct language based on the input ids.

Roberta-Hindi[14]**:** This RoBERTa transformer base model was pre-trained on a large Hindi corpus (a combination of MC4, OSCAR, and indic-nlp datasets) released by flax-community.

MahaRoBERTa[15]**:** A Multilingual RoBERTa (xlm-roberta-base) model fine-tuned on publicly available Marathi monolingual datasets and L3Cube-MahaCorpus.

RoBERTa-Base-Mr[16]**:** The RoBERTa Marathi model was pre-trained on *mr* dataset of C4 (Colossal Clean Crawled Corpus) (Raffel et al., 2019) multilingual dataset.

3.3.3 ALBERT:

A lite BERT for self-supervised learning, Google AI open-sourced ALBERT uses fewer parameters than BERT.

IndicBERT[17]**:** IndicBERT trained on large-scale datasets is a multilingual ALBERT model covering 12 major Indian languages (such as Hindi, Marathi, Bengali, Assamese, English, Gujarati, Oriya, Punjabi, Tamil, Telugu, Kannada and Malayalam) released by Ai4Bharat.

MahaALBERT[18]**:**A Marathi ALBERT model trained on publicly available Marathi monolingual datasets and L3Cube-MahaCorpus.

3.3.4 DistilBERT:

DistilBERT is a small, quick, inexpensive, and light transformer model trained by distilling the BERT base. More than 95% of BERT's performance on the GLUE language understanding benchmark is preserved in this version, which has 40% fewer parameters and runs 60% faster.

mDistilBERT[19]**:** The model is trained on the concatenation of 104 different languages of Wikipedia.

DistilBERTHindi[20]**:** A DistilBERT language model pre-trained on approx. 10 GB of monolingual training corpus, which is taken from OSCAR.

[13] https://huggingface.co/xlm-roberta-base

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

[15] https://huggingface.co/l3cube-pune/marathi-roberta

[16] https://huggingface.co/flax-community/roberta-base-mr

[17] https://huggingface.co/ai4bharat/indic-bert

[18] https://huggingface.co/l3cube-pune/marathi-albert-v2

[19] https://huggingface.co/distilbert-base-multilingual-cased

[20] https://huggingface.co/neuralspace-reverie/indic-transformers-hi-distilbert

4 Result and Analysis

In this section, we discuss the precision, recall and weighted F1 score obtained by training all the variants of BERT on Hindi, Marathi and Bengali datasets. Table 4 represents the results of transformer models trained on the HASOC-Hindi, HASOC-Marathi, and Ben-HateSpeech datasets, where blue, purple, and teal colours indicate multilingual, monolingual, and cross-language models correspondingly. We intentionally prefer a weighted F1 score over an accuracy score to evaluate the models because imbalanced class distribution exists in most classification problems. So, weighted F1 score is a better metric to consider in this scenario. In the training set, the number of sentences for Bengali is double that of Hindi and Marathi. The models trained on the Bengali dataset have better results than Marathi and Hindi. To evaluate our models, we use two class precisions (P_{NOT}, P_{HOF}), recalls (R_{NOT}, R_{HOF}), F1 scores ($F1_{NOT}$, $F1_{HOF}$) then calculate weighted precision (W_P), recall (W_R), and F1 score (W_{F1}) here. At last we calculate $Accuracy$.

$$P_{NOT} = \frac{True_{NOT}}{True_{NOT} + False_{HOF}} \tag{1}$$

$$P_{HOF} = \frac{True_{HOF}}{True_{HOF} + False_{HOF}} \tag{2}$$

$$R_{NOT} = \frac{True_{NOT}}{True_{NOT} + False_{NOT}} \tag{3}$$

$$R_{HOF} = \frac{True_{HOF}}{True_{HOF} + False_{NOT}} \tag{4}$$

$$F1_{NOT} = 2 * \frac{P_{NOT} * R_{NOT}}{P_{NOT} + R_{NOT}} \tag{5}$$

$$F1_{HOF} = 2 * \frac{P_{HOF} * R_{HOF}}{P_{HOF} + R_{HOF}} \tag{6}$$

$$W_P = \frac{P_{NOT} * T_{NOT} + P_{HOF} * T_{HOF}}{T_{NOT} + T_{HOF}} \tag{7}$$

$$W_R = \frac{R_{NOT} * T_{NOT} + R_{HOF} * T_{HOF}}{T_{NOT} + T_{HOF}} \tag{8}$$

$$W_{F1} = \frac{F1_{NOT} * T_{NOT} + F1_{HOF} * T_{HOF}}{T_{NOT} + T_{HOF}} \tag{9}$$

$$Accuracy = \frac{True_{NOT} + True_{HOF}}{T_{NOT} + T_{HOF}} \tag{10}$$

Where $True_{NOT}$ = True-negative (model predicted the texts as NOT, and the actual value of the same is also NOT), $True_{HOF}$ = True-positive (model predicted the texts as HOF, and the actual value of the same is also HOF), $False_{NOT}$ = False-negative (model predicted the texts as NOT, but the true value of the same is HOF), $False_{HOF}$ = False-positive (model predicted the texts as HOF, but the true value of the same is NOT), P_{NOT} = Precision of NOT class, P_{HOF} = Precision of HOF class, R_{NOT} = Recall of NOT class, R_{HOF} = Recall of HOF class, $F1_{NOT}$ = F1 score of NOT class, $F1_{HOF}$ = F1 score of HOF class, T_{NOT} = The total number of NOT class text present in test set, T_{HOF} = The total number of HOF class text present in test set

Best models per datasets: The weighted F1 score for the top four models like MuRILBERT, MahaRoBERTa, NeuralSpaceBERTHi, and XLM-RoBERTa are very close for the Hindi dataset. MahaBERT, MahaRoBERTa, mBERT, and Roberta-Hindi score top for the Marathi dataset. MuRILBERT, BanglaBert, and XLM-RoBERTa models are most suitable for the Bengali dataset. Figure 3 shows the confusion matrix of the best models on three datasets separately.

Monolingual models vs multilingual models: On the Hindi dataset, multilingual models like MuRILBERT and XLM-RoBERTa perform better, but the monolingual model NeuralSpaceBERTHi also gives tough competition. We can conclude that multilingual models perform well, but the difference in performance between monolingual and multilingual models is negligible. MahaBERT and MahaRoBERTa models provide the highest weighted F1 score

Models on HASOC (Hindi)	Precision			Recall			F1 score			Accuracy
	0	**1**	**w.avg.**	**0**	**1**	**w.avg.**	**0**	**1**	**w.avg.**	
mBERT	0.8078	0.7797	0.7949	0.8275	0.8016	0.8156	0.8175	0.7904	0.8050	0.8050
MuRILBERT	0.8695	0.8362	0.8542	0.8266	0.7851	0.8075	0.8475	0.8098	**0.8301**	**0.8308**
NeuralSpaceBERTHi	0.8611	0.8278	0.8458	0.8263	0.7867	0.8081	0.8433	0.8067	0.8264	0.8270
MahaBERT	0.8681	0.8297	0.8504	0.8080	0.7570	0.7845	0.8369	0.7916	0.8161	0.8171
XLM-RoBERTa	0.8218	0.7977	0.8107	**0.8492**	**0.8280**	**0.8394**	0.8352	**0.8125**	0.8247	0.8247
Roberta-Hindi	0.8485	0.8147	0.8329	0.8231	0.7851	0.8056	0.8356	0.7996	0.8190	0.8194
MahaRoBERTa	**0.8892**	**0.8534**	**0.8727**	0.8138	0.7603	0.7892	**0.8498**	0.8041	0.8288	0.8300
RoBERTa-Base-Mr	0.8246	0.7906	0.8089	0.8155	0.7801	0.7992	0.8200	0.7853	0.8040	0.8042
IndicBERT	0.7489	0.7198	0.7355	0.7864	0.7603	0.7744	0.7671	0.7394	0.7543	0.7541
MahaAlBERT	0.8232	0.7913	0.8085	0.8221	0.7900	0.8073	0.8226	0.7906	0.8079	0.8080
mDistilBERT	0.7812	0.7487	0.7662	0.7991	0.7685	0.7800	0.7900	0.7584	0.7754	0.7754
DistilBERTHindi	0.8064	0.7781	0.7934	0.8261	0.8000	0.8141	0.8161	0.7888	0.8035	0.8034
Models on HASOC (Marathi)										
mBERT	0.9019	0.8110	0.8717	**0.9240**	**0.8502**	**0.8995**	0.9128	0.8301	0.8854	0.8848
MuRILBERT	0.8995	0.7878	0.8625	0.8805	0.7536	0.8384	0.8898	0.7703	0.8502	0.8512
NeuralSpaceBERTHi	0.9066	0.8115	0.8751	0.9066	0.8115	0.8751	0.9066	0.8115	0.8751	0.8752
MahaBERT	0.9234	0.8415	0.8962	0.9125	0.8212	0.8822	**0.9179**	**0.8312**	**0.8891**	**0.8896**
XLM-RoBERTa	0.8588	0.7242	0.8142	0.8734	0.7487	0.8320	0.8660	0.7336	0.8221	0.8224
Roberta-Hindi	0.9354	0.8540	0.9084	0.8886	0.7632	0.8470	0.9113	0.8060	0.8764	0.8784
MahaRoBERTa	0.9306	0.8520	0.9045	0.9067	0.8067	0.8735	0.9184	0.8287	0.8886	**0.8896**
RoBERTa-Base-Mr	**0.9688**	**0.8960**	**0.9446**	0.8100	0.5410	0.7209	0.8823	0.6746	0.8135	0.8272
IndicBERT	0.8708	0.6785	0.8071	0.7964	0.5507	0.7150	0.8319	0.6079	0.7577	0.7648
MahaAlBERT	0.9138	0.8095	0.8792	0.8761	0.7391	0.8307	0.8945	0.7726	0.8541	0.8560
mDistilBERT	0.8588	0.6878	0.8021	0.8233	0.6280	0.7586	0.8406	0.6565	0.7796	0.7824
DistilBERTHindi	0.9066	0.7989	0.8709	0.8793	0.7487	0.8360	0.8927	0.7729	0.8530	0.8544
Models on BenHateSpeech										
mBERT	0.9303	0.8630	0.9078	0.9155	0.8362	0.8890	0.9228	0.8493	0.8983	0.8980
MuRILBERT	0.9225	0.8507	0.8985	**0.9406**	**0.8835**	**0.9215**	**0.9314**	**0.8667**	**0.9098**	**0.9095**
XLM-RoBERTa	**0.9463**	**0.8975**	**0.9300**	0.9047	0.8249	0.8781	0.9250	0.8596	0.9032	0.9023
IndicBERT	0.9042	0.8030	0.8704	0.9300	0.8515	0.9038	0.9169	0.8265	0.8867	0.8876
BanglaBERT	0.9333	0.8685	0.9117	0.9207	0.8456	0.8956	0.9269	0.8568	0.9035	0.9033
mDistilBERT	0.8698	0.7290	0.8228	0.9055	0.7941	0.8683	0.8872	0.7601	0.8448	0.8466

Table 4: Precision, Recall, F1 score, and Accuracy of various transformer models on HASOC-Hindi, HASOC-Marathi and BenHateSpeech datasets, respectively

for the Marathi dataset, and mBERT also performs well, whereas MuRILBERT scores a little less. For Bengali, we use only one monolingual pre-trained model, i.e., BanglaBERT; it performs very well, but the MuRILBERT wins marginally. IndicBERT and mDistilBERT models' performance is significantly less on all datasets than in other models. Therefore, developing better resources for the Hindi and Bengali language is necessary as language-specific fine-tuning does not necessarily guarantee the best performance.

Cross-language experiments: During the cross-language experiments, we consider the Marathi models on the Hindi dataset and vice-versa, as both languages share the Devanagari script. MahaRoBERTa performs pretty well on the Hindi dataset, and MahaBERT, RoBERTa-Base-Mr, and MahaALBERT also score sufficiently. NeuralSpaceBERTHi and Roberta-Hindi perform well on the Marathi dataset, but surprisingly, DistilBERTHindi performs poorly on the Hindi dataset rather well on the Marathi dataset.

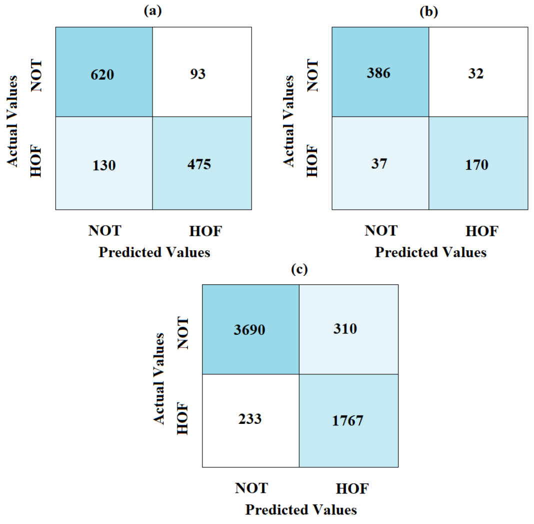


Figure 3: Confusion matrix of Best models such as MuRILBERT for Hindi (a), MahaBERT for Marathi (b) and MuRILBERT for Bengali (c)

5 Conclusion and Future Scope

Significant works are done in English or other major speaking languages. In Indian languages, a very little work has been done, like on Hindi, Bengali, Marathi, Tamil, Malayalam etc. In brief, we conclude this work by: (i) Exploring variants of transformer-based models in Indic languages. (ii) Comparing monolingual and multilingual transformer-based models for hate speech detection data like HASOC-Hindi, HASOC-Marathi and BenHateSpeech. (iii) Cross-language experiments on Hindi and Marathi models. Results show that monolingual training doesn't necessarily ensure superior performance. Multilingual models stood first on Bengali and Hindi datasets, whereas Marathi monolingual models performed the best on Marathi dataset. Our next concentration will be to reduce false-positive and false-negative errors. We also observe that the "0" class precision, recall, and F1 score is slightly higher than the "1" class, indicating the data imbalance. So, in future, techniques like SMOTE (Bowyer et al., 2011), ADASYN (He et al., 2008), or data augmentation (techniques to increase the amount of data) (Nozza, 2022) can be used, which can handle data imbalance. Apart from the technical challenges, the research on hate speech impacts other dimensions, including socio-linguistic issues like freedom of speech and legislation at the national and international levels. Socio-linguistic implications of this research are that some word is used to target a few specific castes, community, colour, etc. A sophisticated hate speech detection system can identify such hated information, restrict its propagation, and alert concerned authorities. In a social context, freedom of speech is related to this issue. So there must be some trade-off between them to maintain peace and harmony.

References

Md. Abul Bashar and Richi Nayak. 2020. Qutnocturnal@hasoc'19: CNN for hate speech and offensive content identification in hindi language. *CoRR*, abs/2008.12448.

Mohit Bhardwaj, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Hostility detection dataset in hindi. *arXiv preprint arXiv:2011.03588*.

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. *CoRR*, abs/2109.13711.

Kevin W. Bowyer, Nitesh V. Chawla, Lawrence O. Hall, and W. Philip Kegelmeyer. 2011. SMOTE: synthetic minority over-sampling technique. *CoRR*, abs/1106.1813.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139.

Anna Glazkova, Michael Kadantsev, and Maksim Glazkov. 2021. Fine-tuning of pre-trained transformers for hate, offensive, and profane content detection in english and marathi. *arXiv preprint arXiv:2110.12687*.

Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. 2008. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328.

Kushal Jain, Adwait Deshpande, Kumar Shridhar, Felix Laumann, and Ayushman Dash. 2020. Indic-transformers: An analysis of transformer language models for indian languages.

Raviraj Joshi. 2022a. L3cube-mahacorpus and mahabert: Marathi monolingual corpus, marathi BERT language models, and resources. *CoRR*, abs/2202.01159.

Raviraj Joshi. 2022b. L3cube-mahacorpus and mahabert: Marathi monolingual corpus, marathi bert language models, and resources. In *Proceedings of The WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 97–101, Marseille, France. European Language Resources Association.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., 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*, pages 4948–4961, Online. Association for Computational Linguistics.

Rajrani Kalra and Ashok K. Dutt. 2019. Exploring linguistic diversity in india: A spatial analysis. *Handbook of the Changing World Language Map*.

Md Rezaul Karim, Bharathi Raja Chakravarthi, John P McCrae, and Michael Cochez. 2020. Classification benchmarks for under-resourced bengali language based on multichannel convolutional-lstm network. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 390–399. IEEE.

Md. Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Sagor Sarker, Mehadi Hasan Menon, Kabir Hossain, Md. Azam Hossain, and Stefan Decker. 2021. Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha P. Talukdar. 2021. Muril: Multilingual representations for indian languages. *CoRR*, abs/2103.10730.

S. Kumari. Nohate at hasoc2020: Multilingual hate speech detection. In *Forum for Information Retrieval Evaluation*, FIRE 2020.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

Zachary Laub. 2019. Hate speech on social media: Global comparisons. *Council on foreign relations*, 7.

Onkar Litake, Maithili Sabane, Parth Patil, Aparna Ranade, and Raviraj Joshi. 2022. Mono vs multilingual bert: A case study in hindi and marathi named entity recognition.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Jouni Luoma and Sampo Pyysalo. 2020. Exploring cross-sentence contexts for named entity recognition with bert. *arXiv preprint arXiv:2006.01563*.

Prasanta Mandal, Apurbalal Senapati, and Amitava Nag. 2022. Hate-speech detection in news articles: In the context of west bengal assembly election 2021. In *Pattern Recognition and Data Analysis with Applications*, pages 247–256, Singapore. Springer Nature Singapore.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In *Forum for Information Retrieval Evaluation*, FIRE 2020, page 29–32, New York, NY, USA. Association for Computing Machinery.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA. Association for Computing Machinery.

JS McCarley, Rishav Chakravarti, and Avirup Sil. 2019. Structured pruning of a bert-based question answering model. *arXiv preprint arXiv:1910.06360*.

Jean-Christophe Mensonides, Pierre-Antoine Jean, Andon Tchechmedjiev, and Sébastien Harispe. 2019. Imt mines ales at hasoc 2019: automatic hate speech detection. In *FIRE 2019-11th Forum for Information Retrieval Evaluation*, volume 2517, pages p–279.

Shubhanshu Mishra and Sudhanshu Mishra. 2019. 3idiots at hasoc 2019: Fine-tuning transformer neural networks for hate speech identification in indo-european languages. In *FIRE (Working Notes)*, pages 208–213.

Sandip Modha, Thomas Mandl, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Tharindu Ranasinghe, and Marcos Zampieri. 2021. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages and conversational hate speech. In *Forum for Information Retrieval Evaluation*, FIRE 2021, page 1–3, New York, NY, USA. Association for Computing Machinery.

Vandan Mujadia, Pruthwik Mishra, and Dipti Misra Sharma. 2019. Iiit-hyderabad at hasoc 2019: Hate speech detection. In *FIRE (Working Notes)*, pages 271–278.

Mayuresh Nene, Kai North, Tharindu Ranasinghe, and Marcos Zampieri. 2021. Transformer models for offensive language identification in marathi. In *FIRE*.

Debora Nozza. 2022. Nozza@LT-EDI-ACL2022: Ensemble modeling for homophobia and transphobia detection. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 258–264, Dublin, Ireland. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Roushan Raj, Shivangi Srivastava, and Sunil Saumya. 2020. Nsit & iiitdwd @ hasoc 2020: Deep learning model for hate-speech identification in indo-european languages. In *FIRE*.

Ananthakrishnan Ramanathan and Durgesh Rao. 2003. A lightweight stemmer for hindi.

Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, Saiful Islam, et al. 2021. Hate speech detection in the bengali language: A dataset and its baseline evaluation. In *Proceedings of International Joint Conference on Advances in Computational Intelligence*, pages 457–468. Springer.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Sagor Sarker. 2020. Banglabert: Bengali mask language model for bengali language understading.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China national conference on Chinese computational linguistics*, pages 194–206. Springer.

Matej Ulčar and Marko Robnik-Šikonja. 2020. Finest bert and crosloengual bert. In *International Conference on Text, Speech, and Dialogue*, pages 104–111. Springer.

Abhishek Velankar, Hrushikesh Patil, Amol Gore, Shubham Salunke, and Raviraj Joshi. 2022. L3cube-mahahate: A tweet-based marathi hate speech detection dataset and BERT models. *CoRR*, abs/2203.13778.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.