Spread Love Not Hate: Undermining the Importance of Hateful Pre-training for Hate Speech Detection

Anonymous Author(s)
Affiliation
Address
email

Abstract

Pre-training large neural language models, such as BERT, has led to impressive gains on many natural language processing (NLP) tasks. Although this method has proven to be effective for many domains, it might not always provide desirable benefits. In this paper we study the effects of hateful pre-training on low resource hate speech classification tasks. While previous studies on English language have emphasized its importance, we aim to to augment their observations with some non-obvious insights. We evaluate different variations of tweet based BERT models pre-trained on hateful, non-hateful and mixed subsets of 40M tweet dataset. This evaluation is carried for Indian languages Hindi and Marathi. This paper is an empirical evidence that hateful pre-training is not the best pre-training option for hate speech detection. We show that pre-training on non-hateful text from target domain provides similar or better results. Further, we introduce HindTweetBERT and MahaTweetBERT, the first publicly available BERT models pre-trained on Hindi and Marathi tweets respectively. We show that they provide state-of-the-art performance on hate speech classification tasks. We also release a gold hate speech evaluation benchmark HateEval-Hi and HateEval-Mr consisting of manually labeled 2000 tweets each.

1 Introduction

The detection of hate speech in social media is a crucial task (Schmidt and Wiegand, 2017). The effect of hateful social media content on mental health of society is still under research, but it is undeniably negative (Kelly et al., 2018; De Choudhury and De, 2014). Twitter is a significant social media platform and is quite popular in India for the past few years. It is used by many politicians, activists, journalists and businessmen as an official medium of communication with the society. Though Article 19 in Indian Constitution guarantees freedom of speech and expression, identifying and curbing hateful tweets is essential to maintain harmony.

Hindi and Marathi are Indo-Aryan languages predominantly spoken India. Both languages are derived from Sanskrit and have 40+ dialects. Marathi being spoken by 83 million people, it is the third largest spoken language in India and tenth in the world. Hindi being spoken by 528 million people is the largest spoken language in India and the third largest in the world.

Identification of hate in NLP has followed a common trend in NLP, the manual feature based classifiers were followed by CNNs and LSTMs which were then superseded by the modern pre-trained transformers (Mullah and Zainon, 2021; Badjatiya et al., 2017; Velankar et al., 2022). The transformer based masked language models (MLM) pre-trained on variety of text data are suitable for general purpose use cases.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

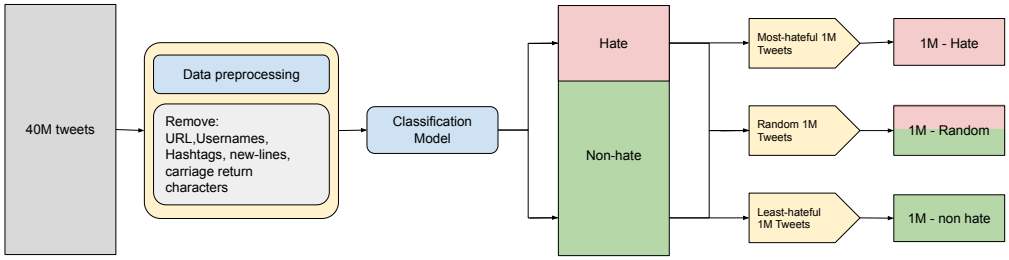

Figure 1: Dataset Generation

Creating a domain-specific bias in pre-training corpus has previously shown state-of-the-art results (Gururangan et al., 2020). Thus, in this paper we try to find the impact of hateful pre-training on hate speech classification. The previous work have shown the positive impact of using Hateful BERT for downstream hate speech identification tasks (Caselli et al., 2020; Sarkar et al., 2021). However, it remains to be verified that the improvements were indeed due to hateful nature of pre-training corpus or is simply a side effect of adaptation to target domain text. The past work in high resource language is thus incomplete and does not provide sufficient evidence to analyze the impact of hateful pre-training. To complete the analysis we pre-train our models using both hateful and non-hateful data from the target domain. Moreover, there is no previous work related to hateful pre-training in low-resource Indic languages. Our work also tries to fill this gap for low resource languages.

While trying to evaluate the impact of pre-training, we build some useful resources for Hindi and Marathi. We introduce two new models MahaTweetBERT and HindTweetBERT pre-trained on 40 million Marathi and Hindi tweets respectively. We use these models along with MuRIL Khanuja et al. (2021), the state of the art Indic multilingual BERT to generate baseline results. To extract the most hateful and least hateful tweets from these 40 million corpora, we classify the tweets using previous state-of-the-art models and choose tweets with highest confidence (most hateful) and lowest confidence (least hateful). We show that the selected data is indeed hateful by randomly choosing 2000 samples each for both the languages and label them manually. To actually see if hateful pre-training has an impact, we compare the performances of models pre-trained on most hateful, least hateful, and random corpora against our baseline on downstream hate speech identification tasks.

2 Related Work

Pre-trained models have obtained remarkable results in many areas of NLP. Although, these pre-trained models are well suited for generalized tasks, they have some limitations in domain-specific tasks. To combat this, numerous domain-specific models have been developed based on the BERT architecture (Devlin et al., 2018). Domain-specific NLP models are pre-trained on in-domain data that is unique to a specific category of text. For example, BioBERT (Lee et al., 2020) is a model that is trained on large-scale biomedical corpora. It outperforms the previous state-of-the-art models on tasks like biomedical named entity recognition and biomedical question answering. Similarly, ClinicBERT (Huang et al., 2019), FinBERT (Yang et al., 2020), LEGAL-BERT (Chalkidis et al., 2020), SciBERT (Beltagy et al., 2019) are models that are pre-trained on clinical notes, financial data, legal documents and scientific text respectively. They show significantly better performance on downstream tasks in their respective domains.

This concept can also be extrapolated for hate speech detection. Models that are pre-trained on exclusively hateful or non-hateful data could work better than models pre-trained on mixed data. For example, HateBERT (Caselli et al., 2020) is a BERT-based model retrained on hateful data extracted from the RAL-E dataset. The RAL-E dataset contains English comments obtained from various subreddits, of which 1.4 million are hateful and are used for pre-training. Similarly, FBERT (Sarkar et al., 2021) is a BERT-based model trained on 1.4 million exclusively hateful tweets from the SOLID dataset (Rosenthal et al., 2020). The SOLID dataset contains 9 million English tweets, of which 1.4 million hateful tweets are used for pre-training. Both these models work better than a vanilla BERT model, when fine-tuned on the downstream training data. Though the results obtained by HateBERT and F-BERT are valid, their experimental setup is not exhaustive and lacks crucial ablations. Given

this, we perform a systematic and more exhaustive study of pre-training on Hindi and Marathi data. We test this on two languages to ensure that the results are not language-specific.

There are various models used for hate speech detection in Marathi. These include monolingual models like MahaBERT (Joshi, 2022c) and multilingual models like MuRIL (Khanuja et al., 2021). Similarly, in Hindi, models like HindBERT (Joshi, 2022a), and MuRIL are the state-of-the-art models for detecting hate speech. We propose a more comprehensive approach than HateBERT and F-BERT, where we pre-train our model on hateful, non-hateful, and random Marathi and Hindi tweets. We then compare the performance of the Marathi models with MuRIL and MahaBERT and the Hindi models with MuRIL and HindBERT to check whether selective pre-training affects model performance.

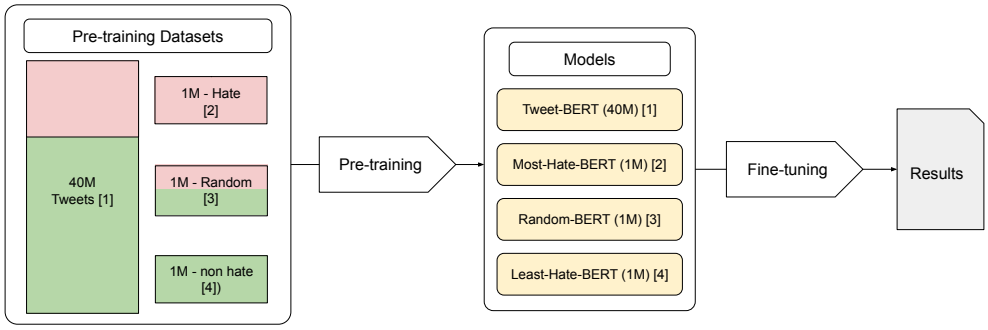

Figure 2: Experimental Setup

3 Dataset description

We create new datasets to observe the effects of using deep learning models trained on primarily hateful data. To test the veracity of our results we have performed experiments with datasets of two different Indian languages, Marathi and Hindi. We follow the same practices and procedures for both of the languages. The datasets consist of a large number of primarily monolingual tweets. From the obtained corpora, we use a threshold value to ensure that tweets contain majority words from the desired language. This ensures that the datasets do not contain tweets with only a few words from the desired language as well as includes the tweets which have a small number of words from other languages (mostly common English words and acronyms such as GST, CAA). All the tweets from both the corpora contain Devanagiri script characters along with numbers. The tweets originally contained usernames, URLs, hashtags and emojis. We clean the data and redact usernames, URLs and hashtags from the datasets to prevent user identification and noise. However, we retain emojis as they might contribute to the semantics of the sentence. All of the tweets are pre-processed before being used for pre-training the models.

3.1 Pre-training corpus

We create a corpus with roughly 40 million tweets each for both languages. All of the other datasets are subsets of these corpora. We illustrate our dataset sampling process in Figure 1. We extract four datasets from the 40M tweet corpus. The primary dataset contains **all of the scraped tweets**.
The second type of dataset contains **1 million tweets randomly sampled** from the primary dataset. The tweets from the 40M datasets were classified into hateful and non-hateful tweets using existing hate speech classification models in the respective languages. For Marathi, the tweets were classified by using the MahaHate-BERT model[1]. MahaHate-BERT is a variant of BERT fine-tuned on the MahaHate dataset (Velankar et al.). For Hindi we use a RoBERTa model (Liu et al., 2019) fine-tuned on Hindi dataset for hate speech classification(Velankar et al., 2021). We record the prediction confidence values along with the predicted labels.
Our third dataset contains **1 million of the most hateful tweets** based on the confidence values of the model. Our fourth dataset consists of the **1 million least hateful tweets** as per the confidence values.

[1]MahaHate-BERT

We choose to use 1 million tweets to maintain consistency with other experiments and also inline with previous work in this area.

3.2 Pre-training corpus verification

To ensure that the models used for segregating hateful and non-hateful tweets are reliable, we manually annotate 2000 tweets of each language and compare them with the predicted labels. We see that the classification models, namely MahaBERT and HASOC-RoBERTa, perform reliably. Concretely, they exhibit a classification accuracy of 77% for Marathi and 75% for Hindi. Since we sample the most hateful and least hateful data for the respective pre-training tasks, we also calculate the classification model accuracy for the most and least hateful predictions in this verification dataset. We report high accuracies in this set- in Marathi, we get 100% and 96.27% accuracy for least and most hateful samples, respectively. Similarly, in Hindi, we get 95.22% and 80.32% accuracy for least and most hateful samples, respectively. The approach can therefore be seen as a credible data selection strategy for pre-training. After manual annotation, we see that out of the 2000 randomly sampled tweets for each language, 803 tweets are hateful in Hindi, and 739 tweets are hateful in Marathi. We plan to release the full verification dataset as a gold benchmark test set in near future.

3.3 Downstream evaluation

We use three datasets for Marathi and two for Hindi to validate and compare our models.

The **HASOC 2021 Marathi dataset** is a dataset presented by the Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages (HASOC 2021) track (Mandl et al., 2021) in the Forum for Information Retrieval Evaluation (FIRE 2021). It contains a total of 2,499 tweets in Marathi manually annotated by native speakers of the language. The dataset was used for a shared task organized by HASOC. The tweets are classified into two categories as hateful and non-hateful tweets.

The **MahaHate two-class dataset** is a hate-speech dataset in Marathi. It contains tweets written in Marathi, annotated manually by four annotators who were fluent in the language. The tweets are classified into two categories, namely hateful or offensive and non-hateful tweets. It contains an equal number of hateful and non-hateful tweets, with a total of 37,500 tweets. The results presented in this paper are based on the models being trained on only 50 percent of the 30,000 tweets of the training dataset. This was primarily done because using the full training data gave similar performance with all the models and the improvements were more prominent in low resource settings.

The **MahaHate four-class dataset** contains Marathi tweets and is similar to the above dataset but contains four classes. The tweets are classified into four categories: hate (HATE), offensive (OFFN), profanity (PRFN), and non-hateful (NOT). The dataset contains a total of 25000 tweets including the train, test and validation datasets. There are equal number of tweets in each of the four categories. Similar to the 2-class dataset, we have used only random half of the training dataset for fine-tuning our models.

We use two datasets for validating and benchmarking our Hindi models. The **HASOC 2021 2-class dataset** contains about 4500 tweets. The classes in the two-class dataset are similar to the HASOC 2021 Marathi dataset.

We also use the **CONSTRAINT 2021 dataset** (Bhardwaj et al., 2020). The original dataset is designed for multilabel classification. We modify the dataset into a 4-class dataset by removing samples which are irrelevant or have multiple ground truth classes. Our modified dataset contains the classes not offensive (NOT), hateful (HATE), defamation (DEF), and offensive (OFF).

Split	Marathi			Hindi	
	HASOC	MahaHate		HASOC	CONSTRAINT
	2-class	2-class	4-class	2-class	4-class
Train	1667	15000	10750	3675	4238
Test	625	3750	2000	919	1217
Validation	207	3750	1500	1532	603

Table 1: Datasets used for training and validation and their respective splits

Dataset	Model	Accuracy	F1 macro	F1 Weighted
Marathi				
MahaHate - 2 class	**MuRIL**	88.57	88.57	88.57
	MahaBERT-V2	89.75	89.75	89.75
	MahaTweetBERT	**89.94**	**89.93**	**89.93**
	mr-most-ht-1m	89.47	89.47	89.47
	mr-random-twt-1m	89.52	89.52	89.52
	mr-least-ht-1m	89.57	89.57	89.57
MahaHate - 4 class	MuRIL	77.83	77.85	77.85
	MahaBERT-V2	79.55	79.55	79.55
	MahaTweetBERT	**79.7**	**79.71**	**79.71**
	mr-most-ht-1m	78.43	78.49	78.49
	mr-random-twt-1m	79.08	79.15	79.15
	mr-least-ht-1m	78.8	78.88	78.88
HASOC	MuRIL	85.7	84.18	85.58
	MahaBERT-V2	88.1	86.76	88.18
	MahaTweetBERT	**89.09**	**87.63**	**89.06**
	mr-most-ht-1m	88.96	87.53	88.95
	mr-random-twt-1m	88.13	86.71	88.19
	mr-least-ht-1m	87.9	86.52	87.98
Hindi				
HASOC 2-class	muril-base-cased	78.59	73.39	77.40
	hindi-bert-v2	80.09	75.99	79.37
	HindTweetBERT	**80.97**	**77.23**	**80.37**
	hi-most-ht-1m	78.36	73.98	77.59
	hi-random-twt-1m	79.73	75.39	78.91
	hi-least-ht-1m	79.44	74.60	78.38
CONSTRAINT 4-class	muril-base-cased	79.39	36.85	74.46
	hindi-bert-v2	81.33	46.49	78.38
	HindTweetBERT	**81.71**	**50.80**	**79.86**
	hi-most-ht-1m	79.74	45.18	77.48
	hi-random-twt-1m	80.92	46.12	78.16
	hi-least-ht-1m	81.55	48.81	79.25

Table 2: Finetuning results on downstream datasets using our models. We report the average scores across five runs with different seed values. The best metrics are shown in bold.

4 Experiments

We conduct extensive experiments to study the effect of different pre-training data compositions on downstream performance. Specifically, we evaluate the different models on three hate detection datasets - MahaHate 4-class, MahaHate 2-class, HASOC 2-class for Marathi. Similarly, we use two Hindi versions of datasets to test our Hindi models - CONSTRAINT 4-class and HASOC 2-class. The number of samples used for training, validation and testing are given in Table 1.

As mentioned in Section 2, Sarkar et al. (2021) and Caselli et al. (2020) present the comparison between retrained transformers and out-of-the-box pre-trained transformers. However, this comparison is substantially inadequate as it does not consider other data compositions for pre-training. Therefore, our ablations include a comprehensive study of models trained on different subsets of the data. As mentioned in Section 3, we experiment with four subsets of the data:

- The complete corpus of 40 million tweets (MahaTweetBERT / HindTweetBERT)
- Most hateful 1 million tweets (mr-most-ht-1m / hi-most-ht-1m)
- Least hateful 1 million tweets (mr-least-ht-1m / hi-least-ht-1m)
- Randomly sampled 1 million tweets (mr-random-twt-1m / hi-random-twt-1m)

We train four models (names as in parenthesis), each corresponding to one subset of the data for each language. We also provide scores on MuRIL (Khanuja et al., 2021) and MahaBERT-V2 (Joshi,

2022b) which we use as baselines of our experiments. In case of Hindi, we use two models, namely HindBERT (Joshi, 2022a) and MuRIL as baselines. The MahaBERT and HindBERT are monolingual BERT models pretrained on publicly available Marathi and Hindi datasets respectively.

For each subset of the data, we train a BERT model using Masked Language Modelling (MLM) technique on the subset. After the pre-training, we individually fine-tune each model on the downstream datasets as shown in Table 1. We finally report the metrics on the test sets of these datasets. Note that we use the same BERT model for all subsets to ensure fairness. We conduct five runs with different seed values for each downstream fine-tuning experiment and report the mean in all cases. Our evaluation pipeline is illustrated in Figure 2. We report our final results in Table 2.

We use the following hyperparameters in our experiments. For MLM training, we train the models for 2 epochs at a learning rate of $2e-5$, with a weight decay of 0.01 and with a mask probability of 0.15. For fine-tuning, we train the models for 25 epochs with a learning rate of $5e-6$ and no weight decay. We use the Hugging Face library (Wolf et al., 2020) for training and hosting our models. All model are uploaded to the HuggingFace platform and their links are available in Appendix A.

5 Results

We hereby analyze the results shown in Table 2. We make some key observations regarding the results as follows.

1. **Hateful BERT is not the best model:** Contrary to our intuition, pre-training only on hateful tweets does not give the best results. In fact, of the three subsets (hateful, non-hateful and random), the model trained on random subset tends to perform better on two of three downstream datasets in Marathi and both downstream datasets in Hindi in terms of macro-F1. Moreover, model trained on non-hateful data performs better than model trained on hateful content for majority of the tasks.
2. **Monolingual retraining shows improvement over multilingual models:** We observe that MuRIL, the multilingual model trained on 17 Indian languages and billions of tokens, has consistently under performed on all datasets. On the other hand, we observe that our models retrained on Hindi tweets outperform MuRIL by a large margin. Note that our Hindi models are essentially MuRIL models retrained on Hindi tweets corpus. The same trend can be seen for Marathi as well. We speculate this is because retraining on sizeable corpora of a particular language augments the multi-lingual pre-training. Specifically, in our case, retraining on languages having medium-sized corpora can outperform cumulative semantic knowledge gained from training on multiple large-sized corpora of different languages, as is in the case of MuRIL.
3. **Model pre-trained on 40 million dataset performs the best on all downstream tasks:** The models MahaTweetBERT and HindTweetBERT outperform all other models on all downstream datasets. This result, although is quite expected, ensures that the large-scale pretraining has indeed helped the model. This provides us with a strong benchmark for future work.

6 Conclusion

In this paper, we test the effect of hateful pre-training on hate speech classification. We pre-train two models, MahaTweetBERT and HindTweetBERT, on 40 million tweets. Additionally, to empirically validate the usefulness of hateful pre-training, we have pre-trained three models on 1 million random, hateful and non-hateful tweets extracted from the aforementioned 40 million tweets for both Marathi and Hindi languages. We compare the performance of these models on standard hate speech detection datasets like HASOC and CONSTRAINT. Our experiments indicate that hateful or non-hateful pre-training does not define the model performance as we observe that models pre-trained on biased tweets do not outperform the ones pre-trained on random tweets. We also observe that our monolingual models fair better than multilingual models like MuRIL. Furthermore, we observe that the models pre-trained on all 40 million tweets perform better than the other, relatively smaller, models. These results are consistent for both Marathi and Hindi languages ensuring that these observations are not language-specific.

References

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*, pages 759–760.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

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

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth international AAAI conference on weblogs and social media*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Raviraj Joshi. 2022a. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages.

Raviraj Joshi. 2022b. L3cube-mahacorpus and mahabert: Marathi monolingual corpus, marathi bert language models, and resources.

Raviraj Joshi. 2022c. L3cube-mahanlp: Marathi natural language processing datasets, models, and library. *arXiv preprint arXiv:2205.14728*.

Yvonne Kelly, Afshin Zilanawala, Cara Booker, and Amanda Sacker. 2018. Social media use and adolescent mental health: Findings from the uk millennium cohort study. *EClinicalMedicine*, 6:59–68.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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. *arXiv preprint arXiv:1907.11692*.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Prasenjit Majumder, Johannes Schaefer, Tharindu Ranasinghe, Marcos Zampieri, Durgesh Nandini, and Amit Kumar Jaiswal. 2021. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages.

Nanlir Sallau Mullah and Wan Mohd Nazmee Wan Zainon. 2021. Advances in machine learning algorithms for hate speech detection in social media: a review. *IEEE Access*.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2020. Solid: A large-scale semi-supervised dataset for offensive language identification. *arXiv preprint arXiv:2004.14454*.

Diptanu Sarkar, Marcos Zampieri, Tharindu Ranasinghe, and Alexander Ororbia. 2021. Fbert: A neural transformer for identifying offensive content. *arXiv preprint arXiv:2109.05074*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

Abhishek Velankar, Hrushikesh Patil, Amol Gore, Shubham Salunke, and Raviraj Joshi. L3cube-mahahate: A tweet-based marathi hate speech detection dataset and bert models.

Abhishek Velankar, Hrushikesh Patil, Amol Gore, Shubham Salunke, and Raviraj Joshi. 2021. Hate and offensive speech detection in hindi and marathi.

Abhishek Velankar, Hrushikesh Patil, and Raviraj Joshi. 2022. Mono vs multilingual bert for hate speech detection and text classification: A case study in marathi. *arXiv preprint arXiv:2204.08669*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*.

A Model links

We present HTTPS URLs for all pretrained models used in our experiments in Table 3.

Marathi	
Model alias	**HTTPS link**
MuRIL	google/muril-base-cased
MahaBERT-V2	marathi-bert-v2
MahaTweetBERT	marathi-tweets-bert
mr-most-ht-1m	mr-most-ht-1m
mr-random-twt-1m	mr-random-twt-1m
mr-least-ht-1m	mr-least-ht-1m
Hindi	
Model alias	**HTTPS link**
MuRIL	google/muril-base-cased
hindi-bert-v2	hindi-bert-v2
HindTweetBERT	hindi-tweets-bert-v2
hi-most-ht-1m	hi-most-ht-1m
hi-random-twt-1m	hi-random-twt-1m
hi-least-ht-1m	hi-least-ht-1m

Table 3: HTTPS links to all pretrained models