L3Cube-MahaHate: A Tweet-based Marathi Hate Speech Detection Dataset and BERT models

Abhishek Velankar[1,3], **Hrushikesh Patil**[1,3], and **Raviraj Joshi**[2,3]
[1]Pune Institute of Computer Technology, Pune
[2]Indian Institute of Technology Madras, Chennai
[3]L3Cube, Pune
{velankarabhishek,hrushi2900}@gmail.com
ravirajoshi@gmail.com

Abstract

Social media platforms are used by a large number of people prominently to express their thoughts and opinions. However, these platforms have contributed to a substantial amount of hateful and abusive content as well. Therefore, it is important to curb the spread of hate speech on these platforms. In India, Marathi is one of the most popular languages used by a wide audience. In this work, we present L3Cube-MahaHate, the first major Hate Speech Dataset in Marathi. The dataset is curated from Twitter and annotated manually. Our dataset consists of over 25000 distinct tweets labeled into four major classes i.e hate, offensive, profane, and not. We present the approaches used for collecting and annotating the data and the challenges faced during the process. Finally, we present baseline classification results using deep learning models based on CNN, LSTM, and Transformers. We explore mono-lingual and multi-lingual variants of BERT like MahaBERT, IndicBERT, mBERT, and xlm-RoBERTa and show that mono-lingual models perform better than their multi-lingual counterparts. The MahaBERT model provides the best results on L3Cube-MahaHate Corpus. The data and models are available at https://github.com/l3cube-pune/MarathiNLP .

Keywords: Natural Language Processing, Convolutional Neural Networks, Long Short Term Memory, FastText, BERT, Hate Speech Detection.

1 Introduction

In the past decade, there has been an expeditious rise in the popularity of online social media platforms all over the globe. People have become more open to sharing their opinions without thinking excessively. This often leads to the spread of hate or offensive speech thereby causing violence and cyberbullying. Hate speech is a kind of abusive language directed towards a community that is underprivileged in terms of race, gender, ethnic origin, disability, etc., or can be an insult or threat to an individual (MacAvaney et al., 2019; Matamoros-Fernández and Farkas, 2021). The users often defy the boundaries of freedom of speech without even realizing it by posting harmful messages and comments (Waseem and Hovy, 2016). Therefore it is today's need to neutralize these activities from proliferating further.

In this work, we consider hate speech detection in the Marathi language, a regional language in India, spoken by over 83 million people across the country (Joshi, 2022). Despite being one of the popular languages in India, work in the area of hate speech detection in Marathi is extremely limited (Mandl et al., 2021; Velankar et al., 2021; Glazkova et al., 2021; Bhatia et al., 2021) as compared to other languages (Del Vigna12 et al., 2017; Romim et al., 2021; Corazza et al., 2020; Schmidt and Wiegand, 2019). Even general text classification in Marathi has received limited attention (Kulkarni et al., 2022, 2021). In this paper, we present, L3Cube-MahaHate Corpus, the largest publicly available hate speech dataset in Marathi. The dataset is collected from Twitter, tagged with four fine-grained labels which are defined as follows-

Hate (HATE): A Twitter post abusing a specific group of people or community based on their religion, race, ethnic origin, gender, geographical location, etc. stimulating violent behaviors.

Proceedings of the 29th International Conference on Computational Linguistic, pages 1–9
October 12–17, 2022.

Offensive (OFFN): A tweet containing harmful language leading to insulting or dehumanizing, at times threatening a particular individual.

Profane (PRFN): A tweet including the use of typical swear words or profane, cursing language which is ordinarily insupportable.

Not (NOT): A post that does not contain any insulting or abusive content or profane words in the language used.

The dataset consists of over 25000 samples tagged manually with the classes explained above. We further provide an extensive study of the data collection approaches, different policies used, and challenges faced during the annotation process as well. We also provide the statistical analysis of our dataset along with the distribution of train, test, and validation data. Lastly, we perform multiple experiments to evaluate state-of-the-art deep learning models on the dataset and provide the baseline results to the community. All the resources will be publicly shared on Github.

The MahaBERT model fined-tuned on L3Cube-MahaHate is termed as MahaHate-BERT[1][2] and is shared publicly on model hub. All the resources are publicly shared on github[3].

2 Related Work

Hate speech detection is considered to be a highly critical problem and a lot of attempts have been made to control it. A significant amount of work can be seen in English text analysis. But recently, efforts have been made towards widening the research in regional languages like Marathi as well.

Gaikwad et al. (2021) presented the Marathi Offensive Language Dataset (MOLD), with nearly 2,500 annotated tweets labeled as offensive and not offensive. It is considered the first dataset for offensive language identification in Marathi. Also, they evaluated the performance of several traditional machine learning models and deep learning models (e.g. LSTM) trained on MOLD.

Bhardwaj et al. (2020) collected over 8200 hostile and non-hostile Hindi text samples from multiple social media platforms like Twitter, Facebook, WhatsApp. Hostile posts were further extended into fake, defamation, hate, and offensive. A total of 8192 posts were collected and tested on various machine learning models using mBERT encoding.

A Hindi-English code-mixed corpus was constructed in Bohra et al. (2018) using the tweets posted online for the duration of five years. Tweets were scrapped using Twitter python API by selecting certain hashtags and keywords from political events, public protests, riots, etc. After removing noisy samples a dataset of 4575 code-mixed tweets was created. The experiments were performed with SVM and Random Forest algorithms along with character and word N-gram features.

In Kulkarni et al. (2021) authors presented a dataset containing over 16000 Marathi tweets, manually tagged in three classes namely positive, negative and neutral. They also provided a policy for tagging sentences by their sentiment. Analysis was performed on CNN, BiLSTM, and BERT models.

Davidson et al. (2017) collected hate phrases identified by Hatebase.org and then used those phrases to collect English tweets from Twitter using Twitter API. The final set of 25k tweets was annotated by CrowdFlower workers with labels hate, offensive and neither. This dataset was then tested on Logistic Regression, Naive Bayes, Decision Trees, random forests, and linear SVMs.

In Geet D'Sa et al. (2021), the authors evaluated the effect of filtering the generated data used for Data Augmentation (DA). This demonstrates up to 7.3% and up to 25% of relative improvements on macro-averaged F1

[1]https://huggingface.co/l3cube-pune/mahahate-bert

[2]https://huggingface.co/l3cube-pune/mahahate-multi-roberta

[3]https://github.com/l3cube-pune/MarathiNLP

on two widely used hate speech corpora.

Ajao et al. (2019) proposed a hypothesis that there exists a relation between fake messages or rumors and sentiments of the texts posted online. The experiments were performed on the standard Twitter fake news dataset and showed good improvement on the same.

Gao and Huang (2018) provided an annotated corpus of hate speech with the context information. This evaluates by using logistic regression and neural network models for hate speech detection around 3% and 4%, and it improves to 7% by combining these two models together.

Mathur et al. (2018) presented MIMCT to detect offensive(Hate or Abusive) Hinglish tweets from the proposed Hinglish Offensive Tweet dataset. Demonstrated the use of the multi-channel CNN-LSTM model for sentiment analysis.

3 Dataset Creation

3.1 Collection

We created the Hate Speech dataset using the tweets posted online by different users across the Maharashtra region considering the period of over the last 5 years. There are plenty of different python libraries available such as Twint[4], GetOldTweets[5], Snscrape[6], etc. which can be used to collect Twitter posts. Twitter provides its own API as well. We used the Twint python library for scraping the tweets.

To obtain the hateful tweets, firstly, we created a list of over 150 bad words in Marathi which are predominantly used by online users to spread hostility. Some of these are typical swear words in Marathi and other offensive words. These words were in Marathi Devanagari script as we are not concerned about Roman or code-mixed text in this work. We will be publishing the final list on GitHub. These words were used as a search query to obtain hate, offensive, and profane tweets. The majority of the tweets that we obtained are related to political and social issues. We also made a note of controversial events with their time frame happening in India in the last couple of years which particularly triggered violence on social media. To avoid bias towards certain words or phrases, we have limited the tweets for a particular search query to a number less than 150. Also, while collecting the tweets, we have not included any reference to the author of the tweet thereby eliminating the bias towards that author.

In our publicly available version of the dataset, we have kept all the hashtags, symbols, emojis, and URLs for anyone to experiment on. However, we have removed all of these while performing the baseline experiments. Furthermore, we will be removing the user mentions from the public dataset to maintain complete user anonymity.

3.2 Annotation

The entire dataset has been labelled manually by the 4 annotators considering four major classes viz. hate, offensive, profane, and not. All the annotators were native Marathi speakers and were fluent in reading and writing in Marathi. The annotation guidelines were set before the tagging exercise. The first 200 sentences were tagged together to further improve the consistency post which sentences were tagged in parallel except for ambiguous sentences. The tweets which were targeted at a single individual thereby criticizing or dehumanizing the individual are tagged as offensive. These tweets were mainly attributed to an individual politician, celebrity, or any random person with the use of singular phrases. The tweets which were targeted at a group of people describing the deficiencies towards race, political opinion, sexual orientation, gender, etc. are tagged as hate. These tweets were majorly concentrated towards political parties or the ruling government. Also, a few samples belong to negative comments on minority groups and gender bias. The tweets which contain swear or profane words are strictly tagged as profane,

[4]https://pypi.org/project/twint/
[5]https://pypi.org/project/GetOldTweets3/
[6]https://github.com/JustAnotherArchivist/snscrape

even if they describe the offensive or hateful category. The tweets that do not satisfy any of the above criteria are simply tagged as NOT. Congratulatory and thanking tweets are tagged as NOT as well. Some sample tweets for the above classes are given in Table 3.

In some cases, the intention of the user behind a tweet cannot be suitably identified. In such cases, the tweets were reviewed again and voting among 4 annotators was used to decide on the labels. Also, we encountered a few tweets where hateful comments were quoted by a news handle. As these posts may indirectly promote violence, we tagged them in the hateful category. To collect the NOT tweets, we selected many Marathi Twitter handles and scraped their tweets, which gave us unbiased data.

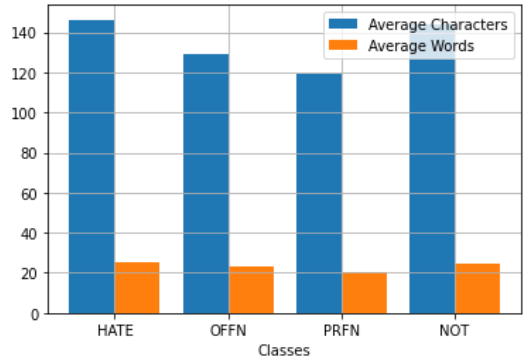


Figure 1: Average characters and words per label

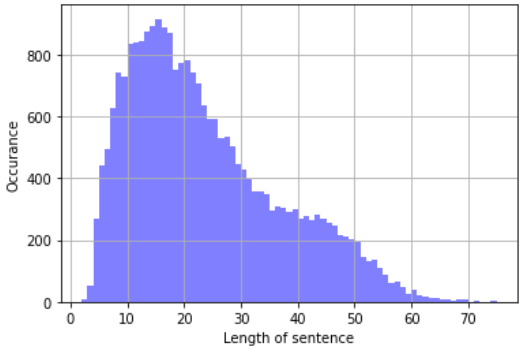


Figure 2: Distribution of the length of a tweet

3.3 Dataset Details

Initially, we collected over 40k tweets in Marathi. Among these, we annotated ~28000 samples. After removing over 3k noisy tweets which particularly included poorly written text i.e. the text with the use of regional words which are not commonly spoken in Marathi or a large number of grammatical mistakes, we randomly selected 6250 samples from each of the 4 classes giving the total count of 25000 tweets. Although this uniform distribution of tweets does not represent the true distribution it makes the model building easier and does not require imbalance handling. We analyzed a few statistics on the dataset. The average number of words per tweet in an entire dataset is 21 and the average number of characters is 113. The label-wise distribution is given in Figure 1. The length of samples varies in the range of 2 to 93. The distribution of the length of tweets and the number of characters per tweet is given in Figures 2 and 3 respectively. The dataset can be used for binary classification as well. To match the number of hateful samples viz. Hate, Offensive, Profane all included, we collected over 12500 extra NOT samples apart from that of 4-class corpus giving an equal distribution of 18750 samples in hateful and non-hateful categories. This binary corpus of 37.5k will also be provided along with the original dataset. The binary dataset is distributed into train, test and validation sets in the ratio of 80:10:10 percent of the total dataset. Table 1 shows the 4-class dataset distribution in training, testing and validation samples.

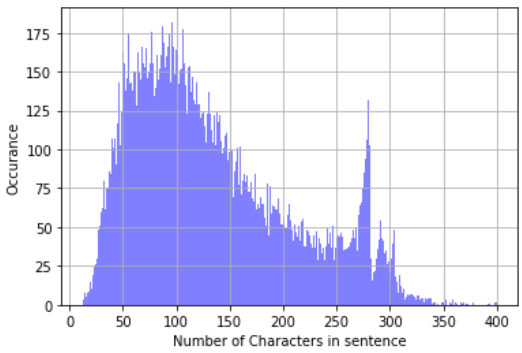


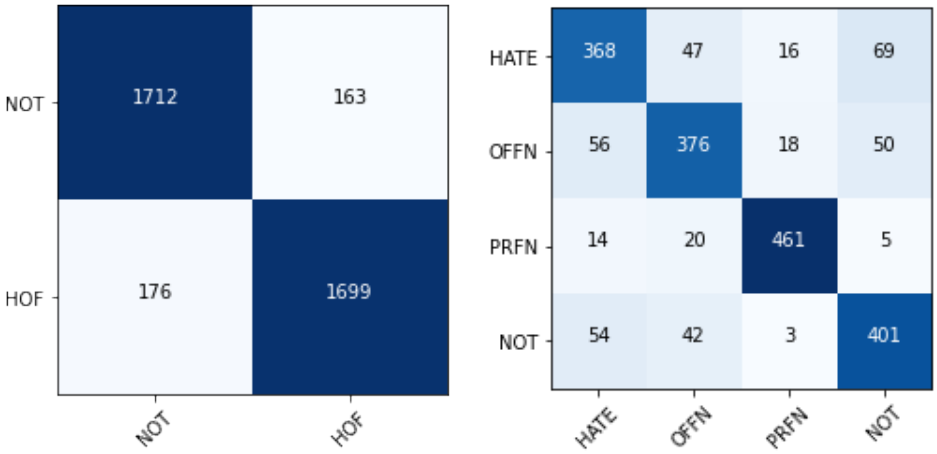
Figure 3: Distribution of the number of characters in a tweet

Split	HATE	OFFN	PRFN	NOT	TOTAL
Train	5375	5375	5375	5375	21500
Test	500	500	500	500	2000
Validation	375	375	375	375	1500

Table 1: Dataset label distribution

Model	Variant	2-Class Accuracy	4-Class Accuracy
CNN	Random	**0.880**	0.703
	Trainable	0.866	0.710
	Non-Trainable	0.870	**0.751**
LSTM	Random	0.857	0.681
	Trainable	0.860	0.691
	Non-Trainable	**0.869**	**0.751**
BiLSTM	Random	0.858	0.699
	Trainable	0.860	0.664
	Non-Trainable	**0.870**	**0.761**
BERT	IndicBERT	0.865	0.711
	mBERT	0.903	0.783
	xlm-RoBERTa	0.894	0.787
	MahaALBERT	0.883	0.764
	MahaBERT	**0.909**	**0.803**
	MahaRoBERTa	0.902	0.803

Table 2: Classification results on different architectures



(a) 2-class classification

(b) 4-class classification

Figure 4: Confusion matrices for the best models

4 Experiments

4.1 Model architectures

We have used multiple state-of-the-art deep learning architectures (Velankar et al., 2021), (Joshi et al., 2021), (Joshi et al., 2019) to obtain the baseline results on 2-class as well as 4-class classification. Before training the models, we have cleaned the data by removing unwanted symbols, user mentions, hashtags. Following algorithms are used for the evaluation of results:

CNN: The CNN model has a 1D convolution layer with a filter of size 300 and a kernel of size 3. It used ReLU activation, followed by max-pooling with pool size 2. the same layers were added again which is

followed by a dense layer of size 50 and ReLU activation. Lastly, the layer with softmax activation and 2 nodes was used. A dropout of 0.3 was used after the 1D max-pooling layer.

LSTM: The LSTM layer with 32 nodes was used. It was followed by a 1D global max-pooling. The dense layer with 16 nodes along with ReLU activation was used, followed by 0.2 dropout. A dense layer with 2 nodes and softmax activation was used as a final layer of the model.

BiLSTM: Bi-LSTM layer with 300 nodes followed by a 1D global max-pooling layer was used. The dense layer was used with 100 nodes and ReLU activation was used with it. This was followed by a dropout of 0.2. At last, the final layer with 2 nodes with activation softmax was used.

BERT: BERT is a bi-directional transformer-based model (Devlin et al., 2019) pre-trained over large textual data to learn language representations. It can be fine-tuned for specific machine learning tasks. We used the following variations of BERT to obtain baseline results:

- Multilingual-BERT (mBERT) - trained on and usable with 104 languages with Wikipedia using a masked language modeling (MLM) objective (Devlin et al., 2018).
- IndicBERT - a multilingual ALBERT model released by Ai4Bharat, trained on large-scale corpora (Kakwani et al., 2020), covering 12 major Indian languages: Assamese, Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil, Telugu.
- XLM-RoBERTa - a multilingual version of RoBERTa (Conneau et al., 2019). It is pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages with the Masked language modeling (MLM) objective and can be used for downstream tasks.
- MahaBERT - a multilingual BERT model (Joshi, 2022) fine-tuned on L3Cube-MahaCorpus and other publicly available Marathi monolingual datasets containing a total of 752M tokens.

4.2 Results

We performed our experiments on CNN, LSTM, and Transformer based models. For CNN and LSTM models, we have used random and fast text initialization for the word embeddings. The pre-trained embeddings were used in both trainable and non-trainable modes. The former means it was used by letting the embedding layer adapt to the training data and the latter by preventing it from being updated during training. Additionally, we used pre-trained language models, particularly the variations of BERT such as IndicBERT, Multilingual BERT, XLM-RoBERTa, and a few custom BERT models to obtain the results. All the 2-class and 4-class accuracies are displayed in Table 2.

In CNN and LSTM based models, non-trainable fast text mode is outperforming other configurations in both the binary and 4-class results. All the monolingual Marathi BERT models are surpassing the multilingual versions of BERT models i.e IndicBERT, mBERT, and xlm-RoBERTa. It was observed that the non-trainable fast text setting for CNN and LSTM based models is performing competitively with the BERT models even surpassing the indicBERT for both classes. The MahaBERT model gives the best binary classification results whereas MahaRoBERTa gives the best 4-class accuracy. The confusion matrices for respective best results are shown in figures 4a and 4b.

5 Conclusion

In this paper, we have presented L3CubeMahaHate - a hate speech dataset containing 25000 distinct samples equally distributed in 4 classes. This is the first major dataset in the domain of hate speech. We also provide the binary version of the dataset of over 37500 samples. We further perform experiments to obtain baseline results on various deep learning models like CNN, LSTM, BiLSTM, and transformer-based BERT models such as IndicBERT,

S.No.	Tweet	English Translation	Tag
1	अशा प्रकारे खोडसाळ बातम्या देणाऱ्या या वृत्तसंस्थाना जोडयाने मारले पाहिजे.	In this way, the news agencies which spread vicious news should be beaten up by the pair of shoes.	HATE
2	स्वतःचे खिसे भरत आहेत. यांना सामान्य जनता मेली तरीही काही फरक पडत न्हाई. स्वार्थी राजकारण नीच वृत्ती ह्या लोकांची.	They are filling their own pockets. Even if the general public dies, it makes no difference to them. Selfish politics, and the mischievous attitude of these people.	HATE
3	काहीही माहिती नसताना दुसर्यांना नालायक म्हणतोस म्हणजे तुझं खुपचं शिक्षण झालं आहे असं वाटते. मुर्खा कुठंही तोंड घालत जाऊ नकोस बेअक्कल.	Calling others incompetent when you don't know anything means you seem to have a lot of education. Idiot, don't put your mouth everywhere, stupid.	OFFN
4	तुझी लायकी काय तू बोलतो कोणा बद्दल काय लाज लज्जा आहे की नाही.	What are your qualifications? Who are you talking about? Do you have any shame or not?	OFFN
5	या मा**द ला वेळीच आवरा नायतर परिणाम भोगायला तयार राहा.	Restraint this m*f*ker on time, otherwise be prepared to suffer the consequences.	PRFN
6	लोकांना असेच चु**या बनवा तुम्ही.. सरसकट आरक्षण काढून टाका आणि सर्वांना जिल्हा परिषद शाळेत शिकवा.	You make people moron like that.. Remove all reservations and teach everyone in Zilla Parishad schools..	PRFN
7	सरकारला आता उत्तर द्यावं लागेल, सामान्य जनतेचा विचार करावा लागेल आता.	The government has to answer now, need to think now of the general public.	NOT
8	तुमचं प्रेम आणि आशीर्वाद यामुळे माझी वाटचाल व्यवस्थित सुरू आहे. अशीच साथ कायम राहू द्या. त्यातूनच मला मातीतल्या माणसांचे प्रश्न, त्यांच्या प्रेरणादायी गोष्टी सांगायचं बळ मिळतं.	Thanks to your love and blessings, my journey is going smoothly. Always keep up this support. It gives me strength to tell the questions of the people of the soil, their inspiring stories.	NOT

Table 3: Sample tweets for each of the 4 classes with English translation .

mBERT and RoBERTa. The dataset is also evaluated on monolingual Marathi BERT models like MahaBERT, MahaALBERT, and MahaRoBERTa. For CNN and LSTM based models, the non-trainable fast text mode outperforms its trainable counterpart in both binary and 4-class classification. In transformer-based models, MahaBERT and MahaRoBERTa give the best results in binary and 4-class classification respectively.

Acknowledgements

This work was done under the L3Cube Pune mentorship program. We would like to express our gratitude towards our mentors at L3Cube for their continuous support and encouragement.

References

Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. 2019. Sentiment aware fake news detection on online social networks. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2507–2511.

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

Mehar Bhatia, Tenzin Singhay Bhotia, Akshat Agarwal, Prakash Ramesh, Shubham Gupta, Kumar Shridhar, Felix Laumann, and Ayushman Dash. 2021. One to rule them all: Towards joint indic language hate speech detection. *arXiv preprint arXiv:2109.13711*.

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of hindi-english code-mixed social media text for hate speech detection. In *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media*, pages 36–41.

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.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–22.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language.

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

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Saurabh Gaikwad, Tharindu Ranasinghe, Marcos Zampieri, and Christopher M. Homan. 2021. Cross-lingual offensive language identification for low resource languages: The case of marathi.

Lei Gao and Ruihong Huang. 2018. Detecting online hate speech using context aware models.

Ashwin Geet D'Sa, Irina Illina, Dominique Fohr, Dietrich Klakow, and Dana Ruiter. 2021. Exploring Conditional Language Model Based Data Augmentation Approaches For Hate Speech Classification. In *TSD 2021 - 24th International Conference on Text, Speech and Dialogue*, Olomouc, Czech Republic.

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*.

Ramchandra Joshi, Purvi Goel, and Raviraj Joshi. 2019. Deep learning for hindi text classification: A comparison. In *International Conference on Intelligent Human Computer Interaction*, pages 94–101. Springer.

Ramchandra Joshi, Rushabh Karnavat, Kaustubh Jirapure, and Ravirai Joshi. 2021. Evaluation of deep learning models for hostility detection in hindi text. *2021 6th International Conference for Convergence in Technology (I2CT)*.

Raviraj Joshi. 2022. L3cube-mahacorpus and mahabert: Marathi monolingual corpus, marathi bert language models, and resources. *arXiv preprint arXiv:2202.01159*.

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 EMNLP*.

Atharva Kulkarni, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, Jayashree Jagdale, and Raviraj Joshi. 2022. Experimental evaluation of deep learning models for marathi text classification. In *Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications*, pages 605–613. Springer.

Atharva Kulkarni, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, and Raviraj Joshi. 2021. L3cubemahasent: A marathi tweet-based sentiment analysis dataset.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

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.

Ariadna Matamoros-Fernández and Johan Farkas. 2021. Racism, hate speech, and social media: A systematic review and critique. *Television & New Media*, 22(2):205–224.

Puneet Mathur, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah. 2018. Did you offend me? classification of offensive tweets in hinglish language. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 138–148.

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.

Anna Schmidt and Michael Wiegand. 2019. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, April 3, 2017, Valencia, Spain*, pages 1–10. Association for Computational Linguistics.

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

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.