# A Twitter BERT Approach for Offensive Language Detection in Marathi

Tanmay Chavan[1,3,†], Shantanu Patankar[1,3,†], Aditya Kane[1,3,†], Omkar Gokhale[1,3,†] and Raviraj Joshi[2,3,†]

[1]*Pune Institute of Computer Technology, Pune*
[2]*Indian Institute of Technology Madras, Chennai*
[3]*L3Cube, Pune*

Abstract

Automated offensive language detection is essential in combating the spread of hate speech, particularly in social media. This paper describes our contribution to the HASOC 2022 Shared Task on Offensive Language Identification in Marathi (Subtask-3A), which handles this crucial task of offensive speech detection in the Marathi Language. In this task, we have to classify a tweet as offensive or non-offensive. We evaluate different mono-lingual and multi-lingual BERT models on this classification task, focusing on BERT models pre-trained with social media datasets. We compare the performance of MuRIL, MahaTweetBERT, MahaTweetBERT-Hateful, and MahaBERT on HASOC 2022 and a combination of HASOC 2021 and HASOC 2022 Marathi datasets. The MahaTweetBERT, a BERT model, pre-trained on Marathi tweets when fine-tuned on the combined dataset (HASOC 2021 + HASOC 2022), outperforms all models with an F1 score of 95.88 on the HASOC 2022 test set.

**Keywords**

Transformers, Hate speech detection, Marathi BERT, Marathi Tweet BERT, HASOC 2022

## 1. Introduction

Offensive speech detection in social media is a crucial task [1]. The impact of cyberbullying and offensive social media content on society's mental health is still under research, but it is undeniably negative [2]. With the increasing number of social media users, offensive speech identification is a crucial task necessary to maintain harmony.

Marathi is an Indo-Aryan language predominantly spoken in the Indian state of Maharashtra. Marathi is a rich language derived from Sanskrit and has 42 dialects. Spoken by 83 million people, it is the third-largest spoken language in India and the tenth in the world.

Our team, Optimize_Prime, participated in the HASOC 2022 [3] shared task on Offensive Language Identification in Marathi, Subtask-3A: Offensive Language Detection. The shared task consisted of classifying tweets as offensive or non-offensive.

In this work, we explore different mono-lingual and multi-lingual pre-trained BERT transformer models for offensive speech detection. Since the evaluation data is based on social media,

*Forum for Information Retrieval Evaluation, December 9-13, 2022, India*

[†]These authors contributed equally.

chavantanamay1402@gmail.com (T. Chavan); shantanupatankar2001@gmail.com (S. Patankar); adityakane1@gmail.com (A. Kane); omkargokhale2001@gmail.com (O. Gokhale); ravirajoshi@gmail.com (R. Joshi)

we focus on Marathi Twitter BERT models and show their superior performance. We evaluate the performance of MuRIL, MahaTweetBERT, MahaTweetBERT-Hateful [4] and MahaBERT on HASOC 2022 and a combination of HASOC 2021 and HASOC 2022 datasets. We provide a detailed analysis of the performance of various mono-lingual and multi-lingual models on the two datasets and reflect on the shortcomings of the models.

## 2. Related Work

Since the advent of social media, detecting offensive language has become an imperative task. Earlier works like Chen et al. [5] used a lexical analysis approach to detect hate speech. Later works like Kumar et al. [6] present a more traditional machine learning-based approach for hate speech detection using feature engineering and models like support vector machines. Aroyehun and Gelbukh [7] uses word embeddings from word2vec, Glove, SSWE, and fastText and uses seven deep learning models, including CNNs, LSTMs, and Bi-LSTMs to detect hate speech from Facebook posts. The introduction of the attention layer and transformers in Vaswani et al. [8] has led to the emergence of various transformer architectures like BERT [9]. These transformers can be fine-tuned on downstream tasks like hate speech to yield exceptional results. There has been some research studying the effect of contextual or domain-specific pre-training. HateBERT [10], and FBERT [11] are BERT models pre-trained on specially curated hate speech data. Both models obtain better results than merely fine-tuning vanilla BERT on a target hate speech dataset.

While much work is available in high-resource languages like English or German, offensive language detection in low-resource languages is relatively less explored. Social media apps' measures to reduce offensive language are generally limited to high-resource languages. Velankar et al. [12] and Gaikwad et al. [13] demonstrate several challenges and limitations faced while performing offensive language detection in Marathi. These limitations warrant a more in-depth study in this particular domain. Multilingual models like MuRIL[14] have been known to perform well on hate speech datasets. However some recent works [15] in Marathi show that Monolingual models like MahaBERT[16] achieve better results than their multilingual counterparts. A much larger corpus consisting of 25000 distinct tweets named L3Cube-MahaHate for hate speech identification for Marathi was proposed in [17]. In our approach, we fine-tune MuRIL, MahaTweetBERT, MahaTweetBERT-Hateful, and MahaBERT on a combination of HASOC 2022 [18], HASOC 2021 [19], and HASOC 2020 data. Both of which are offensive language detection datasets in Marathi.

**Table 1**
Dataset Description for HASOC 2021 and HASOC 2022

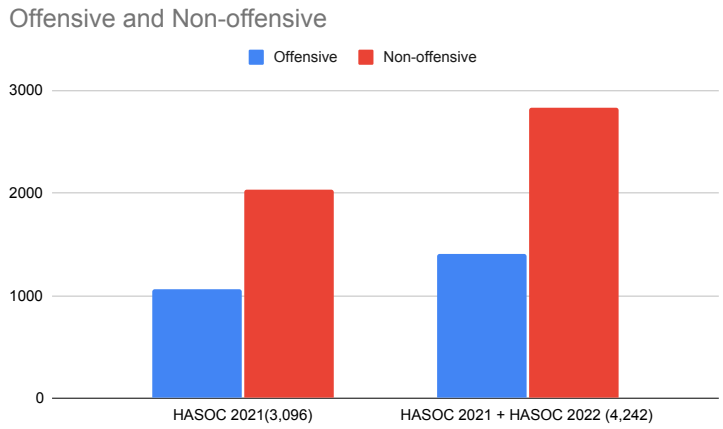
| Datasets | Offensive | Non-offensive | Total |
|---|---|---|---|
| **HASOC 2021** | 1,205 | 669 | 1,874 |
| **HASOC 2022** | 2,034 | 1,062 | 3,096 |

## 3. Dataset

We fine-tune our models on two sets of data. The first dataset consists of the HASOC 2022 training data. The second dataset is a combination of HASOC 2021 and HASOC 2022 data.

### 3.1. HASOC 2022

The HASOC 2022 dataset consists of text from social media. The data points are labeled as offensive and not offensive. We use 70% of the training data to train the model and 30% of the data for validation. The data consists of 3096 data points. Out of these, 2034 are offensive, and 1062 are non-offensive.




### 3.2. HASOC 2021 + HASOC 2022

The HASOC 2021 dataset consists of text obtained from Twitter. The tweets are categorized as offensive and non-offensive. There is a total of 1874 tweets in the dataset. Of these, 1205 are offensive, and 669 are non-offensive. We combine this with the HASOC 2022 data and use the combined dataset for fine-tuning.

## 4. Experiments

We experiment with several models in the course of our experiments. We choose to use BERT-based models as they have shown promising results for text classification. Pre-training these models on large datasets has proven to yield outstanding results on downstream classification tasks in the same language. We describe our methods below.

### 4.1. Data Preprocessing

We preprocessed the data to obtain better results on the classification task. Although the dataset was significantly clean, we performed cleaning operations to ensure the ideal conditions of the data. The provided dataset had redacted username mentions in the tweets and replaced them with the placeholder text '@USER' to protect the identity of the original author of the tweets. We chose to remove the placeholder texts. Our preprocessing methods also cleaned any newline hashtags, URLs, empty parentheses, and newline characters.

### 4.2. Model training

We use the HuggingFace [20] framework for using the models. All the tweets were tokenized by the tokenizer specified by the model before being used by the model. The tokenized text was used by the model backbone. The model's output was then processed through a fully connected feed-forward network layer. We finally used softmax. The models used in this work are described below.

- **MuRIL** is a BERT-based model pre-trained on a large multilingual dataset encompassing several Indian languages.
- **MahaTweetBERT** [1] [4] is a BERT-based model pre-trained on a large monolingual dataset containing tweets written in the Marathi language.
- **MahaTweetBERT-Hateful** [2] is a model which is trained exclusively on hateful Marathi tweets.
- **MahaBERT** [3] [16] is a BERT model pre-trained on a large Marathi dataset containing 725 million tokens.

All of these models are freely available on HuggingFace.

We have used two datasets for training our models. We used the HASOC 2022 dataset and another combined dataset containing samples from the HASOC 2022 dataset and the HASOC 2021 dataset. We performed a split on the HASOC 2022 training dataset to obtain the validation dataset. We recorded the results of our experiments, presented in Table 2. We used the metric of macro F1 score as implemented in the scikit-learn module to remain in line with the official metric. We can see that the MahaTweetBERT model seems to perform very well on both datasets.

We have used the AdamW optimizer with a learning rate of 1e-5 and batch size of 32. We trained the models for 25 epochs. The hyperparameter values and number of epochs remain the same across all the models to maintain consistency between different results.

## 5. Results

We hereby present our results for the HASOC 2022 shared task. We report macro-F1 scores to get a clear idea of the performance of models. Our results on the validation split of the

[1]MahaTweetBERT link: https://huggingface.co/l3cube-pune/marathi-tweets-bert
[2]MahaTweetBERT-Hateful link: https://huggingface.co/l3cube-pune/marathi-tweets-bert-hateful
[3]MahaBERT link: https://huggingface.co/l3cube-pune/marathi-bert-v2

**Table 2**
Our results of all models on validation split of HASOC-22 dataset.

| Model | Macro F1 | |
|---|---|---|
| | HASOC 22 | HASOC 21 + HASOC 22 |
| MuRIL | 88.11 | 95.43 |
| **MahaTweetBERT** | **89.75** | **95.85** |
| MahaTweetBERT-Hateful | 88.60 | 95.28 |
| MahaBERT | 87.92 | 95.62 |

**Table 3**
Our final submission results on testing split of HASOC-22 dataset.

| Model | Dataset | Macro F1 |
|---|---|---|
| **MahaTweetBERT** | HASOC 22 | 91.76 |
| **MahaTweetBERT** | HASOC 21 + HASOC 22 | 95.88 |

HASOC-22 dataset are presented in Table 2. Moreover, our final submissions to the competition with their scores on the test split are shown in Table 3.

We make some key observations from the results. We also spot some interesting patterns that might help in future work.

1. **Combined HASOC dataset performs considerably better than only HASOC-22 dataset:** We see that the combined dataset, HASOC-22, and HASOC-21, when used to train the models, outperforms the performance obtained when trained only on the HASOC-22 dataset. Although expected, this result shows that the language models have not reached their saturation point and can be scaled even further to larger data corpora.
2. **MahaTweetBERT outperforms other models:** MahaTweetBERT is a model pre-trained on a large corpus of Marathi tweets. We observe that this model outperforms all other models in terms of Macro F1. We speculate this is because the downstream dataset, the HASOC-22 dataset, has a high correlation with the pre-training dataset, which is comprised of Marathi Tweets. This shows the importance of domain-specific pre-training in NLProc.
3. **Both models have the same incorrect examples:** The model trained on the combined dataset and the one trained only on the HASOC-22 dataset incorrectly predict the same set of examples. The model trained on the smaller dataset has other incorrectly predicted examples.

## 6. Conclusion

In this paper, we evaluate the performance of various models on the HASOC 2022 and a combined version of the HASOC 2021 and 2022 datasets to observe the ability of our models to detect hate speech in Marathi. We fine-tune models like MuRIL, MahaTweetBERT, and a domain-specific model, MahaTweetBERT-Hateful, pre-trained on 1 million hateful data samples on both datasets. Our experiments show that the models fine-tuned on the combined dataset

perform significantly better. The MahaTweetBERT, pre-trained on 40 million Marathi tweets, outperforms all the other models. We observe that models fine-tuned on both datasets fail to classify some common sentences correctly. In the future, we would like to investigate the reason for this phenomenon.

## 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

[1] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, in: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 1–10. URL: https://aclanthology.org/W17-1101. doi:10.18653/v1/W17-1101.

[2] J. A. Naslund, A. Bondre, J. Torous, K. A. Aschbrenner, Social media and mental health: benefits, risks, and opportunities for research and practice, Journal of technology in behavioral science 5 (2020) 245–257.

[3] S. Satapara, P. Majumder, T. Mandl, S. Modha, H. Madhu, T. Ranasinghe, M. Zampieri, K. North, D. Premasiri, Overview of the HASOC Subtrack at FIRE 2022: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages, in: FIRE 2022: Forum for Information Retrieval Evaluation, Virtual Event, 9th-13th December 2022, ACM, 2022.

[4] S. Patankar, O. Gokhale, A. Kane, T. Chavan, R. Joshi, Spread love not hate: Undermining the importance of hateful pre-training for hate speech detection, arXiv preprint arXiv:2210.04267 (2022).

[5] Y. Chen, Y. Zhou, S. Zhu, H. Xu, Detecting offensive language in social media to protect adolescent online safety, in: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, IEEE, 2012, pp. 71–80.

[6] R. Kumar, A. K. Ojha, S. Malmasi, M. Zampieri, Benchmarking aggression identification in social media, in: Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018), 2018, pp. 1–11.

[7] S. T. Aroyehun, A. Gelbukh, Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling, in: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), 2018, pp. 90–97.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[10] T. Caselli, V. Basile, J. Mitrović, M. Granitzer, Hatebert: Retraining bert for abusive language detection in english, arXiv preprint arXiv:2010.12472 (2020).

[11] D. Sarkar, M. Zampieri, T. Ranasinghe, A. Ororbia, Fbert: A neural transformer for identifying offensive content, arXiv preprint arXiv:2109.05074 (2021).
[12] A. Velankar, H. Patil, R. Joshi, A review of challenges in machine learning based automated hate speech detection, arXiv preprint arXiv:2209.05294 (2022).
[13] S. S. Gaikwad, T. Ranasinghe, M. Zampieri, C. Homan, Cross-lingual offensive language identification for low resource languages: The case of Marathi, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), INCOMA Ltd., Held Online, 2021, pp. 437–443. URL: https://aclanthology.org/2021.ranlp-1.50.
[14] S. Khanuja, D. Bansal, S. Mehtani, S. Khosla, A. Dey, B. Gopalan, D. K. Margam, P. Aggarwal, R. T. Nagipogu, S. Dave, et al., Muril: Multilingual representations for indian languages, arXiv preprint arXiv:2103.10730 (2021).
[15] A. Velankar, H. Patil, R. Joshi, Mono vs multilingual bert for hate speech detection and text classification: A case study in marathi, arXiv preprint arXiv:2204.08669 (2022).
[16] R. Joshi, 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, European Language Resources Association, Marseille, France, 2022, pp. 97–101. URL: https://aclanthology.org/2022.wildre-1.17.
[17] H. Patil, A. Velankar, R. Joshi, L3cube-mahahate: A tweet-based marathi hate speech detection dataset and bert models, in: Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022), 2022, pp. 1–9.
[18] T. Ranasinghe, K. North, D. Premasiri, M. Zampieri, Overview of the HASOC subtrack at FIRE 2022: Offensive Language Identification in Marathi, in: Working Notes of FIRE 2022 - Forum for Information Retrieval Evaluation, CEUR, 2022.
[19] T. Mandl, S. Modha, G. K. Shahi, H. Madhu, S. Satapara, P. Majumder, J. Schäfer, T. Ranasinghe, M. Zampieri, D. Nandini, A. K. Jaiswal, Overview of the HASOC subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.
[20] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.