Marathi Hate Speech Classification

Anonymous ACL submission

Abstract

Hate speech detection is a critical task in natural language processing (NLP) with significant societal impact. In this paper, we focus on hate speech classification in Marathi text, aiming to develop an effective model for identifying hateful, offensive and profane language. We leverage a custom model architecture based on MahaRoBERTa and incorporate additional features to improve classification performance. We manually curate a list of hate, offensive and profane workds in Marathi by looking at various social media posts, etc. Our approach involves tokenizing input sentences and extracting features from these manually curated lists. We trained the model on a balanced dataset. Experiments demonstrate the effectiveness of our model in accurately classifying hate speech in Marathi text. We call our model MaHate.

1 Introduction

Hate speech, defined as speech that attacks a person or group based on attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender, has become an issue in online communication. The rise of social media and online forums has made it easier for individuals to disseminate hateful and discriminatory content, leading to increased concerns about the impact of such speech on society. Detecting and mitigating hate speech online is crucial to maintaining a safe and inclusive online environment.

In this paper, we address the challenge of hate speech detection in Marathi text, a language spoken predominantly in the state of Maharashtra, India. Marathi, like many other languages, faces challenges in detecting hate speech due to the nuances and complexities of language use. Existing approaches to hate speech detection often use a labelled corpus to classify the sentences but do not rely on manually curated lists of offensive and hateful words specific to a given language, hence these approaches may not capture the full range of hate speech present in the language under consideration.

To address this limitation, we propose an approach that combines a custom MahaRoBERTa model (Joshi, 2022) with additional features extracted from manually curated lists of hate, offensive, and profane words in Marathi. By incorporating these features into our model, we aim to improve its ability to detect hate speech and accurately classify text into relevant categories. Through this work, we hope to contribute to the development of effective hate speech detection tools for Marathi and other languages, ultimately fostering a safer and more inclusive online environment for all users.

Our key contributions in this paper are:

- We provide a curated list of popular hateful, offensive, or profane Marathi words to be used as a helper for classification.
- We propose a novel system aware of language-specific derogatory or hateful terms in combination with Hierarchical Attention to perform Hate Speech Detection for Marathi, a low-resource language.
- We conducted extensive experiments for a valid comparison against other state-of-the-art models, including generative AI models.

2 Related Work

Hate speech detection has garnered significant attention in recent years due to its pervasive presence on online platforms. While substantial research exists for English and other major languages, the landscape for low-resource languages like Marathi is relatively small.

Hate Speech Detection in English: English has been the primary focus of hate speech detection research, benefiting from large-scale annotated datasets and advanced language models. Studies have explored various machine learning and deep learning techniques, including Support Vector Machines (SVMs) (Abro et al., 2020), Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) (Fazil et al., 2023). Recent advancements in transformer-based (Vaswani et al., 2017) architectures, such as BERT (Devlin et al., 2019) have been in models like (Mnassri et al., 2022), and have significantly improved performance. Notable datasets like Hate Speech 18 (de Gibert et al., 2018) and the Davidson dataset (Davidson et al., 2017) have served as benchmarks for evaluating models.

Hate Speech Detection in Indian Languages: The development of hate speech detection systems for low-resource languages presents significant challenges. The scarcity of annotated data, coupled with linguistic and cultural complexities, hinders the development of effective models. Some work is done in this field using transformers based models for some low-resource indic languages like Tamil and Telugu (Vegupatti et al., 2024) using Muril (Khanuja et al., 2021), for Assamese language using mBERT, Bangla-BERT (Ghosh et al., 2023).

Hate Speech Models and Datasets in Marathi: Datasets like L3Cube-Mahahate (Patil et al., 2022) with around 25000 Marathi labelled sentences and HASOC 2021 (Modha et al., 2022) have provided a significant contribution to address lack of database for Marathi. These datasets have enabled further research in Marathi hate speech detection, as explored in works like (Nandi et al., 2024; Ghosh and Garain, 2022). A recent paper that attempted to preprocess the data is (Sarode and Sultanova, 2024), they tried combinations of replacing emoticons, and other techniques such as stemming and stop word removal.

3 Methodology

3.1 Data Collection and Preparation

We created three lists of words representing hate speech, offensive language, and profanity in Marathi. The lists were manually curated by native Marathi speakers to properly gauge the category and impact of the words, which might be tough for a standard Language model. The list was created after reviewing hate speech datasets and comments from different social media platforms like Twitter, Instagram and Facebook. The L3-cube-Mahahate dataset (Patil et al., 2022) was used as the base dataset for this project. The text contained many trivial features like Twitter handles, special characters, URLs, etc. All these along with punctuations were removed as a part of preprocessing.

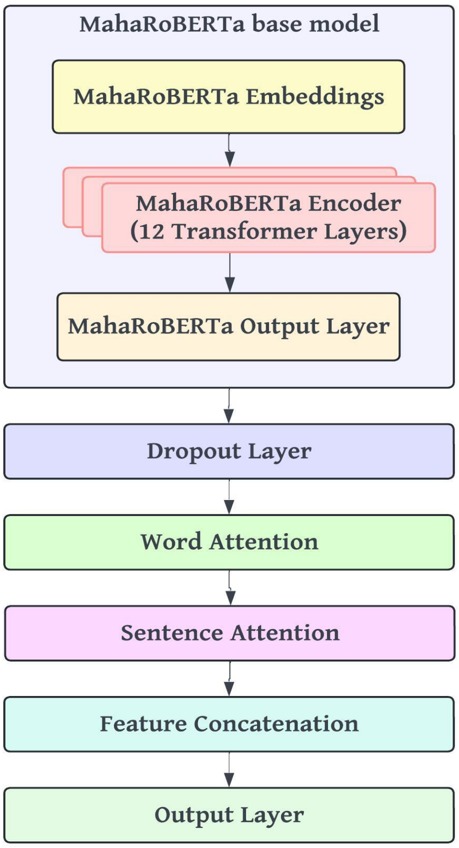

Figure 1: Model Architecture

3.2 Tokenization and Feature Extraction

We used the marathi-roberta tokenizer to tokenize the input sentences. For each input sentence, we extracted additional features representing the presence of words from the hate, offensive, and profanity lists. These features were binary vectors indicating whether each word from the lists was present in the sentence.

3.3 Model Architecture

Our hate speech classification model is based on the MahaRoBERTa architecture, a variant of RoBERTa (Liu et al., 2019) that has been pretrained on Marathi data. We modified the base MahaRoBERTa model to incorporate additional features representing the presence of hate speech, offensive language, and profanity in the input sentences. Figure 1 shows the schematic of the model architecture. The details are explained next.

3.3.1 Input Layer

The input to the model consists of tokenized sentences in Marathi. Each tokenized input sentence is converted into input IDs and an attention mask, which are standard inputs for MahaRoBERTa.

3.3.2 Additional Features

In addition to the tokenized input, our model takes in additional features representing the presence of hate speech, offensive language, and profanity in the input sentences. These additional features are binary vectors of the same length as the word lists for hate speech, offensive language, and profanity. Each element in the vector corresponds to whether the corresponding word from the lists is present in the input sentence.

3.3.3 MahaRoBERTa Base Model

The tokenized input sentences and additional features are passed through the MahaRoBERTa base model, which consists of multiple transformer layers. The layers encode the input sentences into contextualised representations, capturing the semantic meaning of the input text.

3.3.4 Attention layer

We have encountered some inputs containing multiple sentences at once. So, we have implemented a Hierarchical Attention Network inspired from (Yang et al., 2016). This layer calculates which word to focus on in a sentence. Furthermore, a similar process is followed but on sentence level to get a final representation of the given input.

3.3.5 Classifier Layer

The final hidden state output from the MahaRoBERTa base model, along with the additional features, is concatenated and passed through a classifier layer. The classifier layer is a linear layer that maps the concatenated representation to the number of output labels (four in our case: hate speech, offensive language, profanity, and neutral text). The output of the classifier layer is a probability distribution over the four labels, indicating the likelihood of each label for the input sentence.

3.4 Model Training and Fine-Tuning

Our model is initialized using the pretrained MahaRoBERTa architecture, specifically trained on Marathi. The model was trained using AdamW optimizer, with a learning rate of 5e-6 and batch size of 8 which were decided after rigorous testing.

Model	Accuracy
ChatGPT-4o	59.17 %
L3CubeMahahate - xlm-RoBERTa	78.70 %
L3CubeMahahate - MahaRoBERTa	80.30 %
MaHate- xlm-RoBERTa	82.85 %
MaHate- mahaRoBERTa	**83.35 %**

Table 1: Comparison for 4-class classification

Class	Precision	Recall	F1 Score
Hate	73.35 %	79.80 %	76.44 %
Offence	80.89 %	79.60 %	80.24 %
Profane	93.99 %	93.80 %	93.89 %
NOT	84.09 %	78.20 %	81.04 %

Table 2: Class-wise performance metrics

4 Results

4.1 Evaluation

We first compared for 4-class classification against other models and ChatGPT. We assessed our models focusing on configurations that used RoBERTa variants. The MaHate-xlm-RoBERTa, incorporates the XLM-RoBERTa base model with our custom architecture, The MaHate-mahaRoBERTa, utilizes the MahaRoBERTa base model combined with our custom architecture. Along with our models, we tested Mahahate-multi-RoBERTa, MahaRoBERTa, xlm-RoBERTa, mahaBERT models on the same dataset.

For ChatGPT-4o's performance evaluation on hate speech detection, we used the prompt:
"I am providing a Marathi sentence below. Classify it into one of the 4 following classes – 'offensive', 'hate', 'profane', 'not hate'.
`<Sentence>` "
ChatGPT provided predicted classes for each sentence based on this prompt.

The performance of our model is checked on a total of 2000 sentences, with 500 sentences in each class. Table 1 shows the overall results with Table 2 showing classwise details. MaHate-mahaRoBERTA is the best performer. Class-wise analysis shows robust performance across all categories.

Figure 3 shows an example Marathi tweet and how it is classified. The word "nīch" (mean) has the highest influence towards classifiying the tweet as "Hate".

We also studied a 2-class classification problem where all the hate classes are clubbed together. Fig-

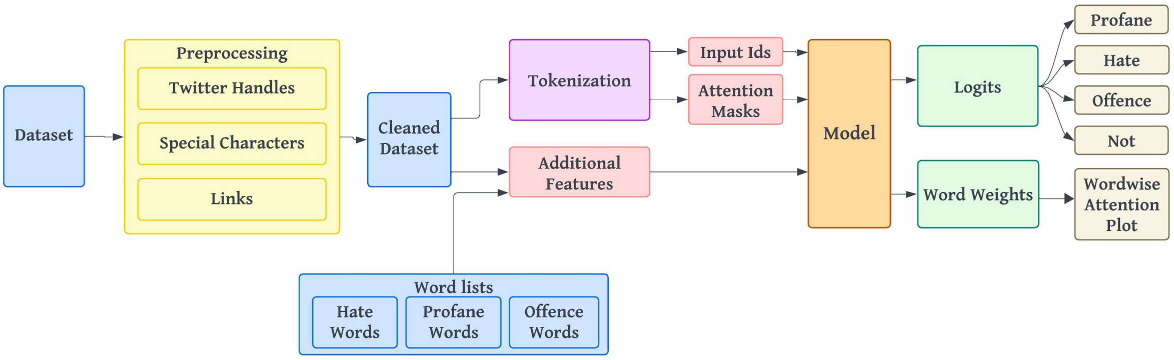

Figure 2: Workflow diagram

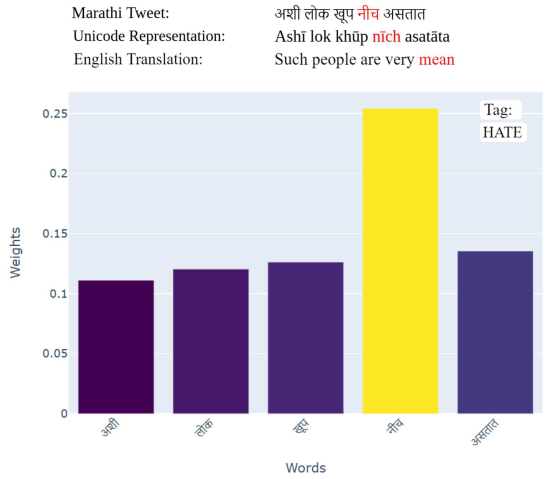
Marathi Tweet: अशी लोक खूप नीच असतात
Unicode Representation: Ashī lok khūp nīch asatāta
English Translation: Such people are very mean

Figure 3: Example of Marathi tweet, predicted label and weights of words in prediction

ure 4 shows the confusion matrix for the same. The accuracy achieved is 91.14%.

4.2 Ablation Study

We analyze the component-wise effect on our model's performance. The model consists of 2 elements, namely, the list of words (L) and hierarchical attention network (HAN). Table 3 shows that both the components are important, although the effect of HAN is greater.

5 Conclusions

In this paper, we proposed a model for classification of Marathi tweets into different categories of hate. We carved a manual list of different kinds of hate words and used them in standard transformer architectures to achieve the best results.

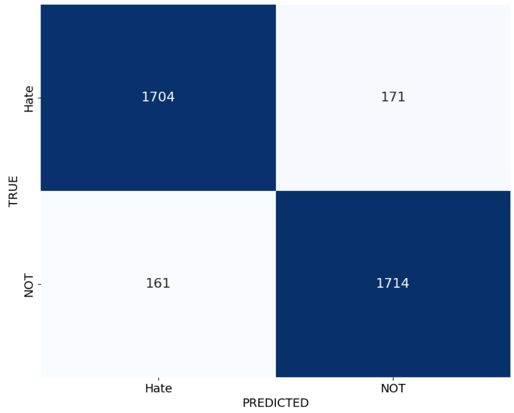

Figure 4: Confusion Matrix for 2 classes

Model	Accuracy
MaHate	83.35%
MaHate– L	82.95%
MaHate– HAN	80.85%
MaHate– L – HAN	80.30%

Table 3: Ablation study

Limitations

We have manually looked at as many tweets as possible for curating the list of different kinds of hate words. While unlikely, it may be still possible that we have missed some important words.

Ethics

We have only used publicly available tweets. While some of them may read vulgar, since the very nature of our work is to identify and classify such tweets, we do not see any other ethical concerns.

References

Sindhu Abro, Sarang Shaikh, Z. Ali, Sajid Ali Khan, Ghulam Mujtaba, and Zahid Hussain Khand. 2020. Automatic hate speech detection using machine learning: A comparative study. *International Journal of Advanced Computer Science and Applications*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*. Association for Computational Linguistics.

Mohd Fazil, Shakir Khan, Bader M. Albahlal, Reemiah Muneer Alotaibi, Tamanna Siddiqui, and Mohd Asif Shah. 2023. Attentional multi-channel convolution with bidirectional lstm cell toward hate speech prediction. *IEEE Access*, pages 16801–16811.

Apurbalal; Ghosh, Koyel; Senapati and Utpal Garain. 2022. Baseline bert models for conversational hate speech detection in code-mixed tweets utilizing data augmentation and offensive language identification in marathi. In *CEUR Workshop Proceedings*, pages 563–574.

Koyel Ghosh, Debarshi Sonowal, Abhilash Basumatary, Bidisha Gogoi, and Apurbalal Senapati. 2023. Transformer-based hate speech detection in assamese. In *2023 IEEE Guwahati Subsection Conference (GCON)*, pages 1–5.

Raviraj Joshi. 2022. 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.

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 Talukdar. 2021. Muril: Multilingual representations for indian languages.

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.

Khouloud Mnassri, Praboda Rajapaksha, Reza Farahbakhsh, and Noel Crespi. 2022. Bert-based ensemble approaches for hate speech detection.

Sandip Modha, Thomas Mandl, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Tharindu Ranasinghe, and Marcos Zampieri. 2022. 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 *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 1–3.

Arpan Nandi, Kamal Sarkar, Arjun Mallick, and Arkadeep De. 2024. Combining multiple pre-trained models for hate speech detection in bengali, marathi, and hindi. *Multimedia Tools and Applications*, pages 1–25.

Hrushikesh Patil, Abhishek Velankar, and Raviraj Joshi. 2022. 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)*, pages 1–9, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Ankur Sarode and Nailya Sultanova. 2024. Detection of hate speech in marathi using language specific preprocessing. *International Journal of Data Science and Advanced Analytics*, 6(6):297–301.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Mani Vegupatti, Prasanna Kumar Kumaresan, Swetha Valli, Kishore Kumar Ponnusamy, Ruba Priyadharshini, and Sajeetha Thavaresan. 2024. Abusive social media comments detection for tamil and telugu. In *Speech and Language Technologies for Low-Resource Languages*, pages 174–187.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.