Hate and Offensive Speech Detection in Hindi and Marathi

Abhishek **Velankar**[1], Hrushikesh **Patil**[1], Amol **Gore**[1], Shubham **Salunke**[1] and Raviraj **Joshi**[2]

[1]*Pune Institute of Computer Technology, Pune, Maharashtra*
[2]*Indian Institute of Technology Madras, Chennai, Tamilnadu*

Abstract

Sentiment analysis is the most basic NLP task to determine the polarity of text data. There has been a significant amount of work in the area of multilingual text as well. Still hate and offensive speech detection faces a challenge due to inadequate availability of data, especially for Indian languages like Hindi and Marathi. In this work, we consider hate and offensive speech detection in Hindi and Marathi texts. The problem is formulated as a text classification task using the state of the art deep learning approaches. We explore different deep learning architectures like CNN, LSTM, and variations of BERT like multilingual BERT, IndicBERT, and monolingual RoBERTa. The basic models based on CNN and LSTM are augmented with fast text word embeddings. We use the HASOC 2021 Hindi and Marathi hate speech datasets to compare these algorithms. The Marathi dataset consists of binary labels and the Hindi dataset consists of binary as well as more-fine grained labels. We show that the transformer-based models perform the best and even the basic models along with FastText embeddings give a competitive performance. Moreover, with normal hyper-parameter tuning, the basic models perform better than BERT-based models on the fine-grained Hindi dataset.

Keywords

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

1. Introduction

Hate speech is often defined as the use of hateful language with the intent of attacking a person or a group to provoke, intimidate, express contempt or cause harm to them or on the basis of their race, religion, ethnic origin, disability or gender [1, 2]. The advancement of technology has led to an increase in the use of social media and its accessibility across the globe. Several online social media users post harmful content without realizing that their actions often offend a person or a group of people [3, 4]. It is therefore important to automatically detect and filter out such harmful content from the massive textual content being posted online every day [5, 6, 7].

Hindi is one of the official languages of India and is spoken by around 45% of its population [8]. Due to its popularity in India, there are a large number of social media activities performed in the Hindi language written in Devanagari script. It is therefore important to detect hate speech in the Hindi language.

Marathi is the native language of Maharashtra state in India. It is spoken by around 83 million people all over the country and it ranks as the third most spoken language in India. People find it easier to express their opinions in regional languages and hence social media activities in Marathi have been quite popular among the Marathi-speaking diaspora. Most of the work in the area of sentiment analysis and hate speech detection has been concentrated on English [9]. Exposure to native low-resource languages has been increasing in recent times. We specifically focus on low-resource Indian languages Hindi and Marathi.

In this work, we treat hate speech detection as a text classification problem and explore various deep learning approaches for the task. The datasets used are provided in the HASOC 2021 shared task [10]. These datasets consisted of text from different Twitter posts, tagged manually as hate and non-hate. The Marathi dataset has binary labels whereas the Hindi dataset consists of binary labels as well as more fine-grained labels namely none, hate, offensive, and profane. We analyze CNN and LSTM based models for the binary classification task. The word embeddings are initialized using corresponding Hindi or Marathi FastText word vectors. We also evaluated transformer-based models, particularly variations of BERT such as indicBERT, mBERT, RoBERTa for Hindi and Marathi [11, 12]. A hierarchical approach is used for the fine-grained 4-class classification task in Hindi where we first distinguish the text between hate and non-hate class and use the text with hate class for further classification into three labels including HATE, OFFN, and PRFN. The hierarchical approach is compared with its direct multi-class counterpart. The best BERT models for each of the tasks are shared publicly[1] [2] [3].

2. Related Work

The low resource nature of Hindi and Marathi languages has limited the extent of work on hate speech detection in these languages. Typical deep learning models like CNN 1D, LSTM, and BiLSTM along with domain-specific word embeddings were evaluated on Hindi-English code mixed dataset in [13]. They also showed that the above deep learning models perform way better than traditional machine learning approaches such as SVM, and random forests.

A comparative study between machine learning and deep learning architectures for hate speech detection is proposed in [14] where datasets containing English tweets have been used. Different combinations of feature engineering have been experimented which include machine learning models like Logistic Regression, Decision Trees, Random Forest, Naive Bayes, etc with TF-IDF and BOW vectorizers. Pre-trained embeddings GLoVe and custom word vectors have been used to train LSTM and GRU models.

In [15] authors compared different machine learning and neural network approaches for hate text speech detection in Hindi, with further classification in hate, offensive, and profane. The

---

[1]https://huggingface.co/l3cube-pune/hate-bert-hasoc-marathi
[2]https://huggingface.co/l3cube-pune/hate-roberta-hasoc-hindi
[3]https://huggingface.co/l3cube-pune/hate-multi-roberta-hasoc-hindi

classical machine learning models like Linear SVM, Adaboost or Adaptive Boosting, Random Forests, Voting Classifier were used and LSTM based deep learning approaches were also used. They observed that machine learning models work better than neural network models in low-resource settings.

Various Hindi text classification approaches have been studied in [8] using BOW, CNN, LSTM, BiLSTM, BERT, and LASER models. The work is particularly focused on Hindi text classification. It is shown that CNN with Hindi fast text embeddings performs the best. Additionally LASER has given very close results to the best performing model as compared to BERT.

In [16] authors propose approaches for hostile post detection in Hindi. Tests were performed on models like CNN, Multi-CNN, BiLSTM, CNN+BiLSTM, IndicBert, mBert along with FastText embedding provided by both IndicNLP and Facebook. This work shows that BERT-based models work slightly better than basic models. The multi CNN model with IndicNLP FastText word embedding performs best within the basic models.

3. Dataset Details

We used the hate speech detection datasets provided in HASOC 2021 shared task for Hindi and Marathi. The text for both datasets was obtained from Twitter. **Marathi Dataset Description [17]:** The dataset consisted of 1874 training samples with an average of 13 words per sentence. The class-wise details are shown in Table 1. It contained a total of 625 testing samples. **Hindi**

Table 1
Marathi dataset label description and distribution

Category	Description	No. of Training Samples
HOF	Hate and Offensive content	669
NOT	Does not contain any hate, offensive, profane content	1205

Dataset Description [18]: The Hindi training dataset included a total of 4594 training samples which were divided into two tasks with an average of 26 words per sentence. Task 1 contained binary labels similar to Marathi i.e. HOF and NOT. Task 2 contained multiclass labels with 4 classes namely NONE, OFFN, HATE, PRFN. Even though labels in task 2 may sound similar, they are different by meaning as described in Table 2. A total of 1532 test samples was provided for both tasks.

4. Model Architectures

We are using common deep learning text classification approaches for the task of Hate speech detection [19]. The models are used directly for binary classification tasks whereas a hierarchical approach is used for multi-labeled fine-grained classification. For each of the models, we selected the epoch giving maximum validation accuracy. We used the learning rate of 0.001 for CNN and LSTM based models whereas 5e-5 for BERT-based models. The general flow of the classification process is outlined in Figure 1 and Figure 2. The models and the configurations

Table 2

Hindi dataset label description and distribution

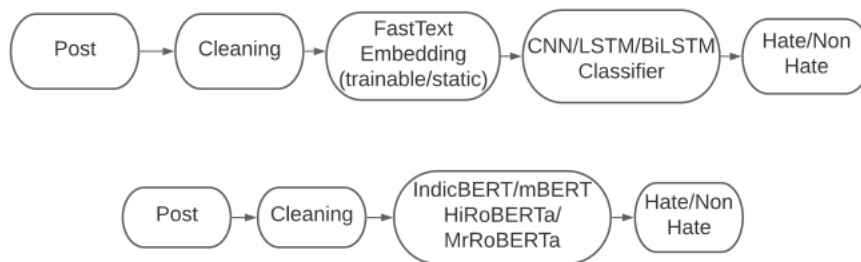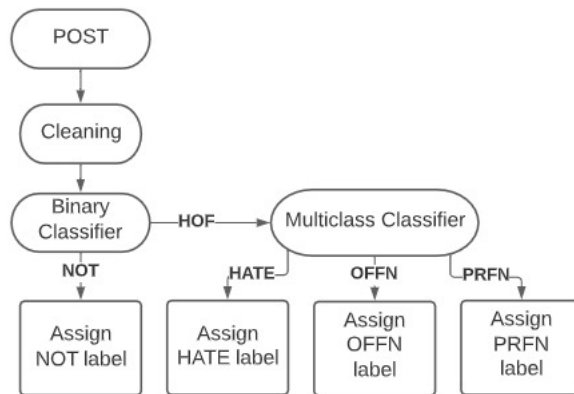| Task 1		
Category	A Description	No. of Training Samples
HOF	Hate and Offensive content	1433
NOT	Does not contain any hate, offensive, profane content	3161
Task 2		
NONE	Does not contain any hate or offensive content	3161
OFFN	The posts contain offensive language	654
HATE	Hate speech content	566
PRFN	Profane words are used	213

Figure 1: Flow of binary classification

Figure 2: Hierarchical approach representation

are outlined below.

CNN: The basic CNN model used a 1D convolution layer with a filter of size 300 and kernel of size 3 with relu activation, followed by max-pooling with pool size 2, the same layers were added again, followed by 1D global max pooling. This is followed by a dense layer of size 50 and relu activation. Finally, the last layer with 2 nodes with softmax activation was used. Dropout of 0.3 was used after the 1D max-pooling layer.

LSTM: For the LSTM model LSTM layer with 32 nodes followed by 1D global max-pooling was used, then a dense layer with 16 nodes along with relu activation was used, followed by 0.2 dropout and finally, a dense layer with 2 nodes with softmax activation was used.

BiLSTM: Bi-LSTM layer with 300 nodes followed by 1D global max-pooling layer was used, then dense layer with 100 nodes along with activation relu was used. This was followed by a dropout of 0.2, then the final layer with 2 nodes with activation softmax was used.

BERT: BERT is a pre-trained language model on a large publicly available text corpus. It is a transformer-based model which is bi-directional in nature. It is pre-trained using two tasks-masked language modeling and next sentence prediction. We evaluated some variations of BERT for both Hindi and Marathi tasks [20].

- Multilingual BERT[4]: Pre-trained on 104 top languages worldwide including Hindi and Marathi.
- IndicBERT[5]: Pre-trained on 12 major Indian languages released by Ai4Bharat.
- roberta-base-mr[6]: Released by flax-community, pre-trained on Marathi with masked language modeling objective.
- roberta-Hindi[7]: RoBERTa base model for Hindi released by flax-community.
- indic-transformers-hi-bert[8]: BERT model pretrained on OSCAR corpus released by neuralspace-reverie.

Hierarchical Approach:

- The first model is trained on task 1 having binary labels HOF and NOT.
- The second model is trained on ternary labels defined in Task 2 by removing entries having NONE values. The ternary labels include OFFN, HATE, and PRFN.
- The test data is passed through the first model to get the corresponding output labels as HOF or NOT.
- The samples predicted as HOF labels are further passed to the second model for classifying them into HATE, OFFN, and PREN labels, results from both models are then combined for the final result.

Table 3
Simple models Evaluation Results

Model	Embedding	Accuracy Score	Macro F1	Macro Precision	Macro Recall
		Marathi			
CNN	Random	0.737	0.721	0.854	0.740
	Trainable FastText	0.841	0.820	0.877	0.818
	Non-Trainable Fastext	0.832	0.817	**0.910**	0.832
LSTM	Random	0.808	0.782	0.854	0.782
	Trainable FastText	0.819	0.795	0.862	0.794
	Non-Trainable Fastext	**0.859**	**0.842**	0.900	0.844
BiLSTM	Random	0.787	0.767	0.868	0.778
	Trainable FastText	0.803	0.777	0.850	0.776
	Non-Trainable Fastext	0.854	0.839	0.909	**0.847**
		Hindi Task 1			
CNN	Random	0.735	0.701	0.80	0.701
	Trainable FastText	0.780	0.740	0.810	0.730
	Non-Trainable Fastext	0.780	**0.760**	**0.850**	**0.760**
LSTM	Random	0.734	0.703	0.808	0.705
	Trainable FastText	0.750	0.710	0.790	0.700
	Non-Trainable Fastext	0.760	0.750	0.830	0.750
BiLSTM	Random	0.751	0.712	0.802	0.708
	Trainable FastText	0.760	0.712	0.781	0.703
	Non-Trainable Fastext	**0.800**	0.745	0.796	0.726
		Hindi Task 2			
		Direct 4 Class			
CNN	Non-Trainable FastText	**0.753**	0.549	**0.618**	0.517
		Hierarchical Approach			
CNN	Non-Trainable FastText	0.734	**0.554**	0.587	**0.534**

5. Results and Discussion

In this work, the performance of different CNN and LSTM based models with and without FastText embeddings was evaluated on HASOC 2021 Marathi and Hindi datasets. Additionally, transformer-based models, particularly variations of BERT were used for comparison. Firstly, all three basic models CNN, LSTM, BiLSTM were trained with random word embedding initialization. The word embeddings were also initialized using pre-trained fast text embedding by IndicNLP and then used in trainable or static mode. The non-trainable fasttext embedding seems more promising than trainable fasttext and random embedding. In this case, the embeddings do not overfit the training data. The results of the basic models are described in Table 3.

Table 4 summarizes the performances of different BERT models. It shows that indic BERT

Table 4
Transformer-based Models Evaluation Results

Model	Accuracy Score	Macro F1	Macro Precision	Macro Recall
Marathi				
indicBERT	**0.880**	**0.869**	**0.871**	**0.867**
mBERT	0.860	0.848	0.844	0.853
RoBERTa-Base-Mr	0.870	0.850	0.858	0.844
Hindi Task 1				
indicBERT	0.770	0.720	0.747	0.708
RoBERTa Hi	**0.800**	**0.763**	**0.778**	**0.754**
Neural space BERT Hi	0.761	0.687	0.746	0.674
Hindi Task 2				
Direct 4 Class				
RoBERTa Hi	0.711	0.521	0.553	0.499
Hierarchical Approach				
RoBERTa Hi	**0.724**	**0.540**	**0.567**	**0.520**

outperforms others in Marathi. For Hindi task 1, the RoBERTa Hindi base model performs the best. For Hindi Task 2, a hierarchical approach is used where two RoBERTa Hindi base models were trained, first for binary and second for ternary classification removing the NONE values. This model performs better than direct multiclass classification but slightly lower than FastText + CNN setting for Task 2. We observe that BERT models are more susceptible to data imbalance in Hindi fine-grained task and requires oversampling from underrepresented classes. Whereas basic models are more robust to such imbalance, the direct 4-way approach performs better than hierarchical classification. The confusion matrices for the best model in each task is shown in Figure 3.
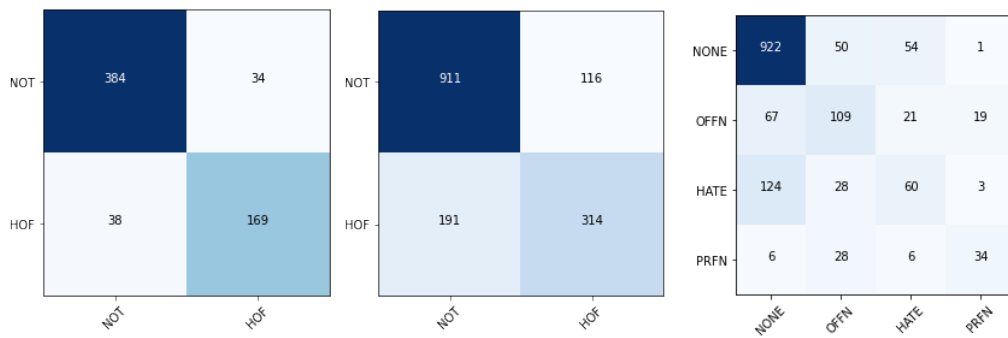

Figure 3: Confusion Matrices for best models in Marathi binary (left), Hindi binary (middle) and Hindi multiclass (right) task respectively

6. Conclusion

In this work, we compared different deep learning approaches on Hindi and Marathi datasets from the HASOC 2021 shared task. The task included both binary and multiclass classification. For binary classification in Marathi and Hindi task 1, CNN and LSTM based models were used along with random and FastText embeddings. Out of these, the LSTM + non-trainable FastText setting worked the best for Marathi. In the case of Hindi, BiLSTM + non-trainable FastText performed better. Additionally, we experimented on different transformer-based BERT models like indicBERT, mBERT, RoBERTa-base for Marathi and RoBERTa base, and Neural space BERT for Hindi. IndicBERT outperformed other models for Marathi whereas RoBERTa performed the best for Hindi. The same RoBERTa model was used for the hierarchical approach. We show that transformer-based models perform better for binary tasks, but even basic models perform competitively. For Hindi task 2, it is shown that CNN + non-trainable FastText model performs slightly better than RoBERTa Hindi model.

Acknowledgments

References

[1] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, O. Frieder, Hate speech detection: Challenges and solutions, PloS one 14 (2019) e0221152.

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

[3] M. Banko, B. MacKeen, L. Ray, A unified taxonomy of harmful content, in: Proceedings of the fourth workshop on online abuse and harms, 2020, pp. 125–137.

[4] J. A. Jiang, M. K. Scheuerman, C. Fiesler, J. R. Brubaker, Understanding international perceptions of the severity of harmful content online, PloS one 16 (2021) e0256762.

[5] 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, 2017, pp. 1–10.

[6] F. Del Vigna12, A. Cimino23, F. Dell'Orletta, M. Petrocchi, M. Tesconi, Hate me, hate me not: Hate speech detection on facebook, in: Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), 2017, pp. 86–95.

[7] S. S. Aluru, B. Mathew, P. Saha, A. Mukherjee, Deep learning models for multilingual hate speech detection, arXiv preprint arXiv:2004.06465 (2020).

[8] R. Joshi, P. Goel, R. Joshi, Deep learning for hindi text classification: A comparison, in: International Conference on Intelligent Human Computer Interaction, Springer, 2019, pp. 94–101.

[9] A. Kulkarni, M. Mandhane, M. Likhitkar, G. Kshirsagar, R. Joshi, L3cubemahasent: A marathi tweet-based sentiment analysis dataset, in: Proceedings of the Eleventh Workshop

on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2021, pp. 213–220.

[10] S. Modha, T. Mandl, G. K. Shahi, H. Madhu, S. Satapara, T. Ranasinghe, M. Zampieri, 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: FIRE 2021: Forum for Information Retrieval Evaluation, Virtual Event, 13th-17th December 2021, ACM, 2021.

[11] 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).

[12] D. Kakwani, A. Kunchukuttan, S. Golla, N. Gokul, A. Bhattacharyya, M. M. Khapra, P. Kumar, inlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, 2020, pp. 4948–4961.

[13] S. Kamble, A. Joshi, Hate speech detection from code-mixed hindi-english tweets using deep learning models, 2018. arXiv:1811.05145.

[14] T. Dhamija, Anjum, R. Katarya, Comparative analysis of machine learning and deep learning algorithms for detection of online hate speech, 2021. arXiv:2108.01063.

[15] V. Mujadia, P. Mishra, D. Sharma, Iiit-hyderabad at hasoc 2019: Hate speech detection, in: FIRE, 2019.

[16] R. Joshi, R. Karnavat, K. Jirapure, R. Joshi, Evaluation of deep learning models for hostility detection in hindi text, in: 2021 6th International Conference for Convergence in Technology (I2CT), IEEE, 2021, pp. 1–5.

[17] S. Gaikwad, T. Ranasinghe, M. Zampieri, C. M. Homan, Cross-lingual offensive language identification for low resource languages: The case of marathi, in: Proceedings of RANLP, 2021.

[18] 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. URL: http://ceur-ws.org/.

[19] A. Kulkarni, M. Mandhane, M. Likhitkar, G. Kshirsagar, J. Jagdale, R. Joshi, Experimental evaluation of deep learning models for marathi text classification, arXiv preprint arXiv:2101.04899 (2021).

[20] T. Wolf, J. Chaumond, L. Debut, V. Sanh, C. Delangue, A. Moi, P. Cistac, M. Funtowicz, J. Davison, S. Shleifer, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 38–45.