Harnessing Pre-Trained Sentence Transformers for Offensive Language Detection in Indian Languages

Ananya Joshi[1,3], Raviraj Joshi[2,3]

[1]MKSSS Cummins College of Engineering for Women, Pune, Maharashtra, India
[2]Indian Institute of Technology Madras, Chennai, Tamil Nadu, India
[3]L3Cube Pune, India

Abstract

In our increasingly interconnected digital world, social media platforms have emerged as powerful channels for the dissemination of hate speech and offensive content. This work delves into the domain of hate speech detection, placing specific emphasis on three low-resource Indian languages: Bengali, Assamese, and Gujarati. The challenge is framed as a text classification task, aimed at discerning whether a tweet contains offensive or non-offensive content. Leveraging the HASOC 2023 datasets, we fine-tuned pre-trained BERT and SBERT models to evaluate their effectiveness in identifying hate speech. Our findings underscore the superiority of monolingual sentence-BERT models, particularly in the Bengali language, where we achieved the highest ranking. However, the performance in Assamese and Gujarati languages signifies ongoing opportunities for enhancement. The goal of our team- 'Sanvadita' is to foster inclusive online spaces by countering hate speech proliferation.

Keywords

Natural Language Processing, Sentence-BERT, Transformers, Hate-speech detection, Offensive language detection, Indian Regional Languages, Low Resource Languages, Text Classification, IndicNLP, BERT

1. Introduction

In today's interconnected world, social media platforms have gained significant influence and have become powerful means of spreading hate speech, often targeting individuals or groups based on factors like race, caste, gender, sexual orientation, or political beliefs. The negative effects of this trend, including cyberbullying and the presence of offensive content, are well-documented and can harm the mental well-being of users. As the number of people using social media continues to grow, it is crucial to develop effective methods to identify and address offensive language to maintain a positive online environment.

Efficient tools for detecting offensive, vulgar, and hateful language on social media platforms are essential because such language can disrupt online discussions and have real-world consequences. This highlights the need for robust Natural Language Processing (NLP) systems capable of effectively recognizing and countering offensive language on these platforms [1].

Our research specifically focuses on the challenge of detecting offensive, profane, and hateful language in low-resource Indian languages, namely Assamese, Bengali, and Gujarati. These languages have received relatively less attention in the field of NLP, and they each have unique

Forum for Information Retrieval Evaluation, December 15-18, 2023, Goa, India
joshiananya20@gmail.com (A. Joshi); ravirajoshi@gmail.com (R. Joshi)

linguistic characteristics that require specialized solutions for addressing offensive content effectively.

Bengali, primarily spoken in West Bengal, India, and Bangladesh, is known for its rich literary tradition and cultural significance. With over 230 million speakers, it ranks as the second most spoken language in India and the seventh in the world. Gujarati, predominantly spoken in the Indian state of Gujarat, contributes significantly to India's linguistic diversity with approximately 55 million speakers. Assamese, spoken primarily in the northeastern Indian state of Assam, is rooted in Sanskrit and includes various dialects, playing a vital role in the linguistic diversity of India's northeastern region.

Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages (HASOC) 2023[1] initiative includes four distinct tasks.
We specifically concentrate on two tasks [2]:

- Task 1B: Identifying Hate, Offensive, and Profane Content in Gujarati
- Task 4: Annihilate Hates - Detecting Hate Speech in Bengali and Assamese

Throughout this paper, we rigorously evaluate the performance of both single-language and multi-language models when applied to the datasets associated with these tasks. We primarily focus on sentence-BERT models for identifying offensive language in social media contexts, showcasing their superior performance. Notably, we present state-of-the-art results on the HASOC 2023 test set using specialized models such as BengaliSBERT, GujaratiSBERT [3], and assamese-bert [4], which have been developed by L3Cube-Pune[2].

2. Related Work

The task of hate speech detection in multilingual contexts has garnered significant attention in recent shared tasks and research endeavors. In this section, we provide an overview of related work, with a focus on studies relevant to our investigation of hate speech detection in low-resource Indian languages.

Several shared tasks have aimed to address hate speech detection challenges. For instance, the paper [5] analyses the systems submitted for the HASOC shared tasks and DravidianLangTech workshop conducted in 2020, focusing on Malayalam, Tamil, and Kannada offensive posts on social media. [6] describes the Subtrack 3 of HASOC-2022, focusing on Offensive Language Identification in Marathi. [7] describes the HASOC 2021 subtask of identification of conversational hate speech in code-mixed languages.

Hindi and Marathi, two prominent Indian languages, have received considerable attention in hate speech detection research. Notable studies include [8, 9, 10, 11, 12], which have contributed to the understanding of hate speech dynamics in these languages. [13] presents a comparative study between monolingual and multilingual BERT models for hate speech detection in Marathi language, while [14] presents a similar comparative analysis with cross-language evaluation for Hindi and Marathi.

[1] https://hasocfire.github.io/hasoc/2023/
[2] https://huggingface.co/l3cube-pune

While Hindi and Marathi have been extensively studied, research efforts have expanded to include languages such as Bengali and Assamese. [15] offers insights into hate speech detection in Bengali, while [16] presents transformer based hate speech detection in Assamese. Similar challenges have been explored in South Indian languages, adding to the linguistic diversity of hate speech research. [17] suggests a weighted ensemble framework to capture hate speech and offensive languages on social platforms posted in code-mixed languages like Hindi–English, Tamil–English, Malayalam–English, Telugu–English, and others. The paper [18] proposes a novel technique of selective translation and transliteration for code-mixed and romanized offensive speech classification in Dravidian languages.

These prior studies provide valuable foundations for our investigation into hate speech detection in low-resource Indian languages, such as Assamese, Bengali, and Gujarati, underscoring the growing recognition of the need to address hate speech in diverse linguistic contexts.

3. Experimental Setup

3.1. Task description

Below, we provide an overview of the tasks:

- **Task 1B: Identifying Hate, offensive and profane content in Gujarati**[3]:
  This task focuses on Hate speech and Offensive language identification for Gujarati. This is a coarse-grained binary classification in a few-shot setting, in which participating systems are required to classify tweets into two classes, namely: Hate and Offensive (HOF) and Non-Hate and offensive (NOT).
  - **(NOT) Non Hate-Offensive** - This post does not contain any hate speech, profane, offensive content.
  - **(HOF) Hate and Offensive** - This post contains hate, offensive, and profane content.
- **Task 4: Annihilate Hates**[4] [19]:
  The objective of the task is to detect hate speech in Bengali, Bodo, and Assamese languages. It is a binary classification task. Each dataset (for the three languages) consists of a list of sentences with their corresponding class (hate or offensive (HOF) or not hate (NOT)). Data is primarily collected from Twitter, Facebook, or youtube comments. Team rank is determined based on the Macro F1 score.

3.2. Datasets

HASOC 2023 provides training datasets tagged as ”NOT” and ”HOF” for binary classification for both Task 1 and Task 4. The main source of data collection is Twitter, Facebook, or YouTube comments. Table 1 shows all dataset statistics. The distribution of offensive and non-offensive tweets in the training dataset of each language is depicted in Figure 1

[3] https://hasocfire.github.io/hasoc/2023/task1.html
[4] https://sites.google.com/view/hasoc-2023-annihilate-hates/home

Table 1
HASOC 2023 datasets statistics for Task 1 and Task 4

		Training			Test
	Label ->	HOF	NOT	Total	Total
Task 1	Gujarati	100	100	200	1196
Task 4	Assamese	2347	1689	4036	1009
	Bengali	515	766	1281	320

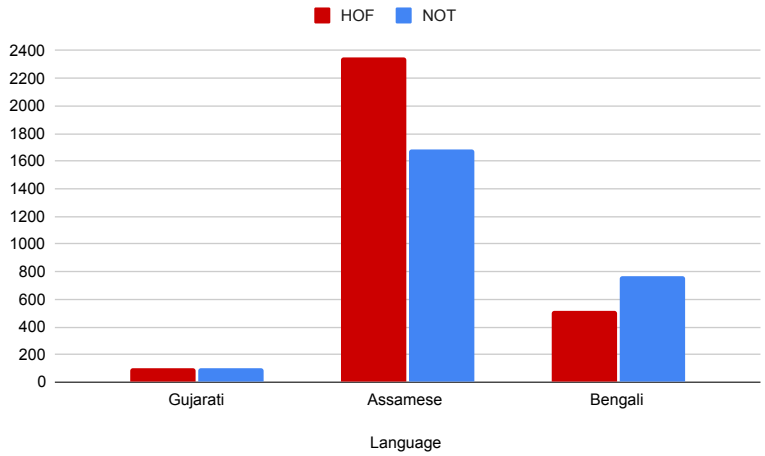

Figure 1: Class distribution of tweets in HASOC 2023 training datasets for Task 1 and Task 4

3.3. Preprocessing

In order to enhance the accuracy of our classification task, we conducted data preprocessing to improve the data quality. We engaged in cleaning procedures to optimize the data conditions, which included eliminating punctuation marks, URLs, usernames, handles, hashtags, numbers, and Roman characters. Additionally, our preprocessing methods addressed issues such as newline characters, excessive spaces, and empty parentheses. Notably, we made a deliberate decision to retain emojis, as they contribute significantly to conveying the sentiment of the text and were observed to yield superior results.

Label encoding: We encode Class label into a unique number for each task: ”HOF” to ”1”, and ”NOT” to ”0”

3.4. Models and Training Setup

BERT [20] models are pre-trained on a massive corpus of text data, where they learn to predict masked words within sentences. Then, they are fine-tuned on specific downstream tasks using labeled data. Sentence-BERT (SBERT) [21] models are trained by learning fixed-size embeddings for sentences using siamese or triplet network architectures that aim to optimize similarity

scores between related sentences and minimize distances between them in embedding space.

While BERT focuses on word-level representations, SBERT models are designed to capture the semantic meaning of sentences, including subtle nuances and context, by producing fixed-size sentence embeddings. Hate speech often relies on the overall context and phrasing of a sentence, making SBERT's sentence-level understanding more relevant. Hate speech classification often requires an understanding of context and context-dependent variations in meaning. SBERT models, leverage contextual information by considering the surrounding words in a sentence, making them more adept at recognizing the intended sentiment or tone. Traditional BERT models, while powerful, may struggle to understand the nuances of entire sentences and their emotional or hateful intent.

The papers [22, 3] show that the Sentence-BERT models outperform the corresponding BERT variants in understanding context-specific information. Hence, we primarily utilize the monolingual and multilingual SBERT models for Gujarati and Bengali languages. The Assamese language, however, lacks quality datasets and powerful models such as Sentence-BERT. Hence, we use the monolingual assamese-bert and the multilingual indic-bert model.

- For Task 1, we use the pre-trained monolingual model GujaratiSBERT[5] and the multilingual IndicSBERT[6] model.
- For Task 4, we use pre-trained monolingual models of BengaliSBERT[7], bengali-bert[8], assamese-bert[9] and the multilingual models IndicSBERT, indic-bert[10].

For both tasks, we initialize a classification model using the BERT architecture and freeze the first six layers of the model. Next, we train the model using the provided training data for 4 epochs with the default learning rate.

4. Results

We conducted training on a range of models using the complete training dataset and subsequently employed these models to predict classes for the provided test dataset. In all the tasks, the texts are classified into 2 categories- HOF, indicating the presence of hateful content, or NOT- indicating no offensive content. The outcomes are presented in Table 2, and the evaluation metric employed for determining the team's leaderboard ranking was the Macro F1 Score. We have included all the task results in accordance with the leaderboard presentation. Additionally, we explored the efficacy of multiple pre-trained BERT and SBERT models but submitted only the most successful run for evaluation, omitting the submission of other runs due to their subpar performance.

We achieved the **top ranking (rank 1)** among 21 participating teams for Task 4- Bengali language, because of the highest Macro F1 Score obtained using the BengaliSBERT model. The

[5] https://huggingface.co/l3cube-pune/gujarati-sentence-bert-nli
[6] https://huggingface.co/l3cube-pune/indic-sentence-bert-nli
[7] https://huggingface.co/l3cube-pune/bengali-sentence-bert-nli
[8] https://huggingface.co/l3cube-pune/bengali-bert
[9] https://huggingface.co/l3cube-pune/assamese-bert
[10] https://huggingface.co/ai4bharat/indic-bert

BengaliSBERT model outperforms the bengali-bert and other multilingual models like MuRil, Indic-bert and IndicSBERT. For Task 4- Assamese language, we attained rank 6 among 20 teams through the use of assamese-bert model. For Task1- Gujarati, we stand at Rank 10 among 17 participating teams. The best score was given by GujaratiSBERT model, outperforming the gujarati-bert and other multilingual models like MuRil, Indic-bert and IndicSBERT.

Table 2
Macro F1 scores obtained from various models, along with the Ranks achieved in Task1 and Task4 of HASOC 2023

Task	Language	Model	MACRO F1	Rank
Task 1	Gujarati	**GujaratiSBERT**	**0.7324**	10
		IndicSBERT	0.7291	
Task 4	Assamese	**assamese-bert**	**0.7065**	6
		indic-bert	0.6788	
Task 4	Bengali	**BengaliSBERT**	**0.7703**	**1**
		IndicSBERT	0.7409	
		indic-bert	0.7121	

5. Conclusion

Through this paper, we describe our approach for hate and offensive speech detection in three Indian languages. We utilize the HASOC 2023 datasets for fine-tuning the pretrained BERT and SBERT models and testing their performance. Our findings reveal that monolingual Sentence-BERT models consistently outperform both monolingual BERT models and multilingual counterparts in the realm of hate speech identification. Notably, we secured the highest ranking for Bengali language, while the lower rankings in Assamese and Gujarati languages underscore the ongoing need for enhancements in these domains. Looking ahead, we are committed to exploring various strategies to elevate the performance of Assamese and Gujarati models. Our overarching goal is to contribute to the advancement of more inclusive and comprehensive tools for combatting online hate speech, ultimately fostering online spaces characterized by tolerance and respect.

Acknowledgments

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. Velankar, H. Patil, R. Joshi, A review of challenges in machine learning based automated hate speech detection, arXiv preprint arXiv:2209.05294 (2022).
[2] T. Ranasinghe, K. Ghosh, A. S. Pal, A. Senapati, A. E. Dmonte, M. Zampieri, S. Modha, S. Satapara, Overview of the HASOC subtracks at FIRE 2023: Hate speech and offensive

content identification in assamese, bengali, bodo, gujarati and sinhala, in: Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE 2023, Goa, India. December 15-18, 2023, ACM, 2023.
[3] S. Deode, J. Gadre, A. Kajale, A. Joshi, R. Joshi, L3cube-indicsbert: A simple approach for learning cross-lingual sentence representations using multilingual bert, arXiv preprint arXiv:2304.11434 (2023).
[4] R. Joshi, L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages, arXiv preprint arXiv:2211.11418 (2022).
[5] B. R. Chakravarthi, D. Chinnappa, R. Priyadharshini, A. K. Madasamy, S. Sivanesan, S. C. Navaneethakrishnan, S. Thavareesan, D. Vadivel, R. Ponnusamy, P. K. Kumaresan, Developing successful shared tasks on offensive language identification for dravidian languages, 2021. arXiv:2111.03375.
[6] T. Ranasinghe, K. North, D. Premasiri, M. Zampieri, Overview of the hasoc subtrack at fire 2022: Offensive language identification in marathi, 2022. arXiv:2211.10163.
[7] S. Satapara, S. Modha, T. Mandl, H. Madhu, P. Majumder, Overview of the hasoc subtrack at fire 2021: Conversational hate speech detection in code-mixed language, Working Notes of FIRE (2021) 13–17.
[8] A. Velankar, H. Patil, A. Gore, S. Salunke, R. Joshi, Hate and offensive speech detection in hindi and marathi, arXiv preprint arXiv:2110.12200 (2021).
[9] T. Chavan, S. Patankar, A. Kane, O. Gokhale, R. Joshi, A twitter bert approach for offensive language detection in marathi, arXiv preprint arXiv:2212.10039 (2022).
[10] 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.
[11] K. Ghosh, A. Senapati, U. Garain, Baseline bert models for conversational hate speech detection in code-mixed tweets utilizing data augmentation and offensive language identification in marathi, in: Fire, 2022. URL: https://api.semanticscholar.org/CorpusID:259123570.
[12] S. Ghosal, A. Jain, Hatecircle and unsupervised hate speech detection incorporating emotion and contextual semantics, ACM Trans. Asian Low-Resour. Lang. Inf. Process. 22 (2023). URL: https://doi.org/10.1145/3576913. doi:10.1145/3576913.
[13] A. Velankar, H. Patil, R. Joshi, Mono vs multilingual bert for hate speech detection and text classification: A case study in marathi, in: IAPR Workshop on Artificial Neural Networks in Pattern Recognition, Springer, 2022, pp. 121–128.
[14] K. Ghosh, D. A. Senapati, Hate speech detection: a comparison of mono and multilingual transformer model with cross-language evaluation, in: Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation, De La Salle University, Manila, Philippines, 2022, pp. 853–865. URL: https://aclanthology.org/2022.paclic-1.94.
[15] M. Das, S. Banerjee, P. Saha, A. Mukherjee, Hate speech and offensive language detection in bengali, arXiv preprint arXiv:2210.03479 (2022).
[16] K. Ghosh, D. Sonowal, A. Basumatary, B. Gogoi, A. Senapati, Transformer-based hate speech detection in assamese, in: 2023 IEEE Guwahati Subsection Conference (GCON), 2023, pp. 1–5. doi:10.1109/GCON58516.2023.10183497.
[17] P. K. Roy, S. Bhawal, C. N. Subalalitha, Hate speech and offensive language detection in dravidian languages using deep ensemble framework, Computer Speech & Language 75

(2022) 101386. URL: https://www.sciencedirect.com/science/article/pii/S0885230822000250. doi:https://doi.org/10.1016/j.csl.2022.101386.

[18] S. Sai, Y. Sharma, Towards offensive language identification for dravidian languages, in: Proceedings of the first workshop on speech and language technologies for Dravidian languages, 2021, pp. 18–27.

[19] K. Ghosh, A. Senapati, A. S. Pal, Annihilate Hates (Task 4, HASOC 2023): Hate Speech Detection in Assamese, Bengali, and Bodo languages, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.

[20] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[21] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).

[22] A. Joshi, A. Kajale, J. Gadre, S. Deode, R. Joshi, L3cube-mahasbert and hindsbert: Sentence bert models and benchmarking bert sentence representations for hindi and marathi, in: Science and Information Conference, Springer, 2023, pp. 1184–1199.