CUET_INSights@NLU of Devanagari Script Languages: Leveraging Transformer-based Models for Target Identification in Hate Speech

Farjana Alam Tofa, Lorin Tasnim Zeba, Md Osama and Ashim Dey
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u1904008, u1904043, u1804039}@student.cuet.ac.bd, ashim.cse.cuet@gmail.com

Abstract

Hate speech detection in multilingual content is a challenging problem especially when it comes to understanding the specific targets of hateful expressions. Identifying the targets of hate speech whether directed at individuals, organizations or communities is crucial for effective content moderation and understanding the context. A shared task on hate speech detection in Devanagari Script Languages organized by CHIPSAL@COLING 2025 allowed us to address the challenge of identifying the target of hate speech in the Devanagari Script Language. For this task, we experimented with various machine learning (ML) and deep learning (DL) models including Logistic Regression, Decision Trees, Random Forest, SVM, CNN, LSTM, BiLSTM, and transformer-based models like MiniLM, m-BERT, and Indic-BERT. Our experiments demonstrated that Indic-BERT achieved the highest F1-score of 0.69, ranked 3rd in the shared task. This research contributes to advancing the field of hate speech detection and natural language processing in low-resource languages.

1 Introduction

Hate speech promotes hostility and discrimination toward certain people or groups and creates major challenges in preserving social harmony. Detecting and identifying hate speech is especially complex in multilingual contexts, where harmful messages may target specific groups. This process is important to better understand the intent and impact of harmful language. The "Shared Task on Natural Language Understanding of Devanagari Script Languages" at CHIPSAL@COLING 2025 aimed to address this challenge, particularly through Subtask C which focused on identifying hate speech targets in Devanagari-scripted text. The goal was to classify the targets of hate speech: "individual," "organization," or "community". Their workshop paper (Sarveswaran et al., 2025) offered us an opportunity to engage with these challenges in processing South Asian languages and to advance our work on hate speech detection and target identification in this context. The proposed approach can be used in content moderation systems to help platforms detect and reduce hate speech in different low-resourced languages. It can assist policymakers by providing a reliable method to track and analyze online hate speech.

In our participation, we explored different models to identify hate speech targets and tried to solve this problem with two significant contributions.

- Investigated the effectiveness of several ML, DL, and transformer models for identifying hate speech targets and examining the errors to obtain important insights about the detection procedure.
- In particular, leveraged the transformer-based Indic-BERT model which has proven effective for the particular use case in Devanagari script languages.

This study shows how advanced models like transformers can improve hate speech detection, and target identification supporting better language understanding.

2 Related Work

The difficulty of identifying hate speech and abusive language has prompted numerous research using a range of languages and methodologies. Recent advancements have focused on addressing hate speech in low-resource languages. The CHUNAV dataset offers a valuable resource for analyzing hate speech in Hindi during elections, capturing nuanced socio-political themes (Jafri et al., 2024). Similarly, the IEHate dataset provides insights into political hate speech in Hindi, highlighting the benefits of human and automated meth-

Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025), pages 267–272
January 19, 2025.

ods in this domain (Jafri et al., 2023). For Nepali, NEHATE facilitates hate speech analysis in local election discourse, contributing to inclusive online dialogue (Thapa et al., 2023). NAET introduces anti-establishment discourse in Nepali, covering unique aspects like hate speech to enhance political sentiment analysis (Rauniyar et al., 2023). Additionally, the Karaka model provides foundational resources for Bhojpuri, aiding NLP development in this language (Ojha, 2019). For Marathi, L3CubeMahaSent offers a structured sentiment analysis dataset, filling a gap for Indian languages (Kulkarni et al., 2021). Itihasa, a large-scale Sanskrit translation dataset highlights the complexity of ancient texts and challenges current translation models (Aralikatte et al., 2021). Hate speech research has addressed diverse forms of toxic content including racism, sexism, and religious bias, while also discussing challenges in real-world applications (Parihar et al., 2021). A review of hate speech detection methods revealed inconsistent results and limited dataset reliability (Alkomah and Ma, 2022). CNN, LSTM, and BERT models proved effective for hate speech detection in Hindi and Marathi and simpler architectures also performed competitively when augmented with FastText embeddings (Velankar et al., 2021). The Dravidian shared task for Malayalam showed m-BERT's strong performance. It highlights the transformer model's potential in misinformation detection for low-resource languages (Osama et al., 2024). An evaluation dataset, HateCheckHIn, was developed to address the challenges of multilingual hate speech detection, focusing on error analysis and diagnostic insights, particularly for Hindi (Das et al., 2022). In Tamil, a study focusing on caste and migration-related hate speech found that M-BERT was highly effective. It highlights the model's suitability for handling nuanced social contexts in low-resource settings (Alam et al., 2024).

3 Task and Dataset Description

With the rise of social media, hate speech has become a significant issue often targeting specific groups. This shared task (Thapa et al., 2025) focuses on hate speech detection in languages using the Devanagari script. It identifies the target of hate speech in a given sentence, classifying it as either "individual," "organization," or "community." The dataset for this task consists of hate speech texts in Devanagari script covering languages such as Nepali, Marathi, Sanskrit, Bhojpuri, and Hindi. This dataset is organized to support accurate classification of hate speech targets as outlined below:

Individual: Hate speech aimed at a specific person.

Organization: Hate speech targeting institutions or groups.

Community: Hate speech directed at larger communities.

Here, Table 1 provides the distribution of samples across training, validation, and test sets. The dataset is imbalanced, with the Community class having the fewest samples (406 texts), compared to Individual (1,534 texts) and Organization (1,223 texts).

Classes	Train	Valid	Test
Individual	1,074	230	230
Organization	856	183	184
Community	284	61	61
Total	2,214	474	475

Table 1: Dataset distribution.

4 Methodology

The methods and approaches employed to address the issue raised in the preceding part are briefly summarized in this section. Through careful analysis, our research recommends utilizing a transformer-based model employing Indic-BERT (Kakwani et al., 2020). Figure 1 provides a concise visualization of our methodology, outlining the key steps involved in our approach.

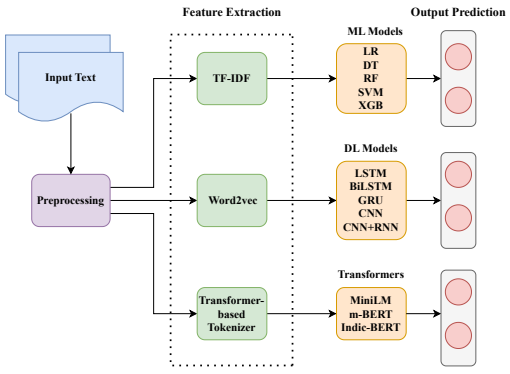

Figure 1: An abstract view of our methodology

4.1 Preprocessing

We translated Bhojpuri tweets into Hindi to ensure uniformity and enhance compatibility with multilingual language models. Basic preprocessing

steps, such as removing special characters, stopwords and empty spaces were also applied to clean the text.

4.2 Feature Extraction

To capture meaningful features for different model types, three feature extraction techniques are employed. For machine learning models, the Term Frequency-Inverse Document Frequency (TF-IDF) (Qaiser and Ali, 2018) approach is used. For deep learning models, word embeddings are generated using the Word2Vec (Ma and Zhang, 2015) technique. And transformer models use architecture-compatible tokenizers for tokenization.

4.3 Model Building

In our research, we explored a variety of ML, DL and transformer-based models.

4.3.1 ML models

We trained traditional ML models such as Logistic Regression, Decision Trees, Random Forest, Support Vector Machines and Extreme Gradient Boosting on TF-IDF features. These models identify patterns statistically but may struggle with the complexity of contextual and linguistic nuances in hate speech.

4.3.2 DL models

The deep learning models include LSTM (Sherstinsky, 2020), BiLSTM (Xu et al., 2019), GRU (Dey and Salem, 2017), CNN (Alzubaidi et al., 2021) and a hybrid CNN+RNN model. These models capture semantic linkages in tweets by using Word2Vec embeddings. Each DL model was trained for 10 epochs with a batch size of 32.

4.3.3 Transformer-based models

The transformer-based models include MiniLM (Wang et al., 2020), m-BERT (Yu et al., 2024) and Indic-BERT (Kakwani et al., 2020). These models are fine-tuned using transformer-specific tokenizers to handle multilingual text efficiently. Transformers outperform ML and DL models because they process entire sentences using attention mechanisms, capturing context and long-range dependencies. They also benefit from pre-training on large multilingual corpora and handle complex scripts like Devanagari with better precision, which reduces information loss.

5 Results & Discussion

In this section, we provide comparisons of the performance achieved by different machine learning, deep learning, and transformer-based methods. The performance evaluation of various classifiers for the targets of hate speech identification showcases valuable details about how well they can predict. We also fine-tuned particularly m-BERT and Indic-BERT by adjusting learning rates, batch sizes, and epochs with the fixed Adam optimizer and Sparse Categorical Cross-Entropy (CCE) loss function in Table 2.

Hyperparameters	**m-BERT**		**Indic-BERT**	
Optimizer	Adam	Adam	Adam	Adam
Loss Function	Sparse CCE	Sparse CCE	Sparse CCE	Sparse CCE
Learning rate	5e-05	3e-05	2e-05	1e-05
Epochs	12	10	8	5
Batch size	8	16	8	8

Table 2: Summary of tuned hyper-parameters

By modifying these hyper-parameters, we tried to improve the model's performance across all metrics. We observed that increasing the number of epochs improved accuracy, with models reaching nearly very high at the end. m-BERT was trained for 12 epochs due to steady improvement, while Indic-BERT was trained for fewer epochs (5-8) due to faster convergence. A summary of the precision (P), recall (R), and macro-F1 (MF1) scores for each model on the test set is presented in Table 3. Among ML models, LR performed best with an MF1 score of 0.60. For DL models, CNN+RNN achieved the highest MF1 score of 0.62. Transformer models outperformed both ML and DL, with m-BERT achieving an MF1 of 0.68, while Indic-BERT emerged as the best overall with an MF1 of 0.69.

Classifier	**P**	**R**	**MF1**
LR	0.61	0.64	0.60
DT	0.55	0.56	0.55
RF	0.59	0.63	0.59
SVM	0.55	0.62	0.57
XGB	0.59	0.61	0.58
LSTM	0.57	0.58	0.57
BiLSTM	0.57	0.57	0.57
GRU	0.56	0.57	0.57
CNN	0.61	0.63	0.61
CNN + RNN	0.62	0.63	0.62
MiniLM	0.67	0.66	0.66
m-BERT	0.70	0.69	0.68
Indic-BERT	**0.74**	**0.67**	**0.69**

Table 3: Results of various models on the test dataset.

5.1 Quantitative Discussion

The results underscore the effectiveness of transformer-based architectures, particularly Indic-BERT, in detecting the target of hate speech. By incorporating contextual embeddings from pre-trained language models like Indic-BERT, our classification system achieves enhanced accuracy. Indic-BERT performs better as it supports four Devanagari languages (Nepali, Marathi, Hindi, Sanskrit) while m-BERT supports only three (Nepali, Marathi, and Hindi), provides broader coverage of Devanagari-script languages compared to m-BERT. To address the class imbalance issue, we applied class weights during training to give more importance to the minor Community class. Additionally, data augmentation techniques such as generating synthetic examples or paraphrasing could further improve the representation of the 'Community' class. The confusion matrix in Figure 2 provides a detailed breakdown of our model's performance.

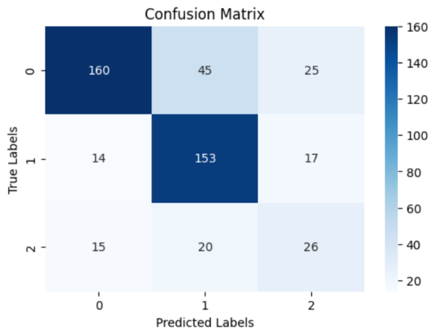

Figure 2: Confusion matrix of our best performing model

The model correctly classifies 160 label-0, 153 label-1 and 26 label-2 instances. However, it misclassifies 45 samples of label-0 as label-1 and 25 as label-2. Label-2 shows significant errors, with 15 misclassified as label-0 and 20 as label-1.

5.2 Qualitative Discussion

Figure 3 showcases some sample predictions made by our Indic-BERT model. Among these, samples 1, 4, and 5 are correctly classified, indicating the model's ability to handle diverse linguistic constructs in Devanagari script. However, certain instances reveal challenges for the model. For example, sample 2 which discusses a cricket match in the Devanagari script context, is misclassified as label 1 instead of the correct label 2. Sample 3 related to political content is also misclassified, suggesting potential difficulties in distinguishing political expressions within the Devanagari script.

Test Sample	Actual	Predicted
Sample 1: दिपायल सिलगढी नगर प्रमुखमा तीन जनाको उम्मेदवारी https://t.co/FhOrzCkPXI (Dipayal Silgarhi Nagar Mayor three candidates https://t.co/FhOrzCkPXI)	2	2
Sample 2: दार्चुलामा भारतीयले बम बिस्फोट गराउँदा नेपाली मारिए, कालिकोटमा नेपालीले राखेको बम बिस्फोट हुँदा २ जना मारिए । यो बम राखिने दुई कारण नेपाली कांग्रेस र माओवादी अहिले घाँटी जोडेर भोट माग्दैछन् है ।#NoNotAgain (A Nepali was killed when the Indians detonated a bomb in Darchula, 2 people were killed when a bomb planted by Nepali exploded in Calikot. The two reasons for keeping this bomb are that the Nepali Congress and the Maoists are now clamoring for votes.#NoNotAgain)	2	1
Sample 3 : एमाले सँग मिलेर पचास बर्ष शासन गर्न कांग्रेसलाई सरापेको प्रचण्ड अचेल कांग्रेससँग मिलेर रेनुलाई जिताईदिएमा दश पन्ध्र बिस बर्ष सम्म शासन गर्न प्रतिबद्ध रहनेछु रे। (Prachanda, who cursed Congress to rule together with UML for fifty years, now if Renu is defeated along with Congress, I will be committed to rule for ten fifteen twenty years.)	0	1
Sample 4 : आखिरकार यह तीनों लुटेरी पाटिया हाथ मिला ही ली यह सब मिलकर पंजाब को लूटने की कोशिश कर रहे हैं पंजाब की जनता से अनुरोध है कि इन सब पार्टियों से बच कर रहे यह सब पार्टियां मिलकर अरविंद केजरीवाल जी की इमानदार पार्टी को हराना चाहते हैं #PunjabElections2022 https://t.co/31zYHWRL1D (Ultimately, these three robber gangs have joined hands and together they are trying to loot Punjab. It is a request to the people of Punjab that all these parties, who are avoiding these parties, together want to defeat Arvind Kejriwal ji's honest party. #PunjabElections2022 https://t.co/31zYHWRL1D)	1	1
Sample 5 : @narendramodi @SherBDeuba फर्किने बेला त्यो देउवा लाई पनि सङ्गै लानुहोला। 🙏 (When @narendramodi @SherBDeuba returns, take that Deuba with you. 🙏)	0	0

Figure 3: Examples of the Indic-BERT model's anticipated outputs with English translations

6 Conclusion

Our study on hate speech target identification in Devanagari-script languages demonstrates that transformer-based models, particularly IndicBERT, achieved the highest F1 score of 0.69, outperforming both machine learning and deep learning models. Despite challenges with low-resource languages like Bhojpuri, tailored preprocessing and feature extraction techniques provided valuable insights. Although this work focuses on Devanagari-script languages, the methodology can be adapted to other low resource scripts by using pre-trained models like m-BERT, MiniLM or Indic-BERT which work well with multilingual data. Future advancements in language-specific models could further improve hate speech detection in diverse multilingual contexts. Additionally, exploring multimodal approaches may significantly improve the accuracy and robustness of hate speech identification.

Limitations

A primary limitation of this study lies in converting Bhojpuri text into Hindi through a manually created vocabulary as the models employed lack training in the Bhojpuri language. This vocabulary-based conversion may not fully capture all nuances and context in Bhojpuri which could lead to potential inaccuracies in the model's performance. Additionally, the lack of high-quality annotated data for low-resource languages limits the robustness of the models. Future exploration of methods with more annotated samples in low-resource languages like Bhojpuri would enhance model accuracy and generalizability. Developing multilingual embeddings or pre-trained transformer models specifically for dialects in the Devanagari script would also address limitations in vocabulary conversion.

References

Md Alam, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. Cuet_nlp_manning@ lt-edi 2024: Transformer-based approach on caste and migration hate speech detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 238–243.

Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Information*, 13(6):273.

Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. 2021. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8:1–74.

Rahul Aralikatte, Miryam De Lhoneux, Anoop Kunchukuttan, and Anders Søgaard. 2021. Itihasa: A large-scale corpus for sanskrit to english translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 191–197.

Mithun Das, Punyajoy Saha, Binny Mathew, and Animesh Mukherjee. 2022. Hatecheckhin: Evaluating hindi hate speech detection models. *arXiv preprint arXiv:2205.00328*.

Rahul Dey and Fathi M Salem. 2017. Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, pages 1597–1600. IEEE.

Farhan Ahmad Jafri, Kritesh Rauniyar, Surendrabikram Thapa, Mohammad Aman Siddiqui, Matloob Khushi, and Usman Naseem. 2024. Chunav: Analyzing hindi hate speech and targeted groups in indian election discourse. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Farhan Ahmad Jafri, Mohammad Aman Siddiqui, Surendrabikram Thapa, Kritesh Rauniyar, Usman Naseem, and Imran Razzak. 2023. Uncovering political hate speech during indian election campaign: A new low-resource dataset and baselines.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, 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 the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

Atharva Kulkarni, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, and Raviraj Joshi. 2021. L3cubemahasent: A marathi tweet-based sentiment analysis dataset. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 213–220.

Long Ma and Yanqing Zhang. 2015. Using word2vec to process big text data. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 2895–2897. IEEE.

Atul Kr Ojha. 2019. English-bhojpuri smt system: Insights from the karaka model. *arXiv preprint arXiv:1905.02239*.

Md Osama, Kawsar Ahmed, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024. Cuet_nlp_goodfellows@ dravidianlangtech eacl2024: A transformer-based approach for detecting fake news in dravidian languages. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 187–192.

Anil Singh Parihar, Surendrabikram Thapa, and Sushruti Mishra. 2021. Hate speech detection using natural language processing: Applications and challenges. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1302–1308. IEEE.

Shahzad Qaiser and Ramsha Ali. 2018. Text mining: use of tf-idf to examine the relevance of words to documents. *International Journal of Computer Applications*, 181(1):25–29.

Kritesh Rauniyar, Sweta Poudel, Shuvam Shiwakoti, Surendrabikram Thapa, Junaid Rashid, Jungeun Kim, Muhammad Imran, and Usman Naseem. 2023. Multi-aspect annotation and analysis of nepali tweets on anti-establishment election discourse. *IEEE Access*.

Kengatharaiyer Sarveswaran, Bal Krishna Bal, Surendrabikram Thapa, Ashwini Vaidya, and Sana Shams. 2025. A brief overview of the first workshop on challenges in processing south asian languages (chipsal). In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*.

Alex Sherstinsky. 2020. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306.

Surendrabikram Thapa, Kritesh Rauniyar, Farhan Ahmad Jafri, Surabhi Adhikari, Kengatharaiyer Sarveswaran, Bal Krishna Bal, Hariram Veeramani, and Usman Naseem. 2025. Natural language understanding of devanagari script languages: Language identification, hate speech and its target detection. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*.

Surendrabikram Thapa, Kritesh Rauniyar, Shuvam Shiwakoti, Sweta Poudel, Usman Naseem, and Mehwish Nasim. 2023. Nehate: Large-scale annotated data shedding light on hate speech in nepali local election discourse. In *ECAI 2023*, pages 2346–2353. IOS Press.

Abhishek Velankar, Hrushikesh Patil, Amol Gore, Shubham Salunke, and Raviraj Joshi. 2021. Hate and offensive speech detection in hindi and marathi. *arXiv preprint arXiv:2110.12200*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Guixian Xu, Yueting Meng, Xiaoyu Qiu, Ziheng Yu, and Xu Wu. 2019. Sentiment analysis of comment texts based on bilstm. *Ieee Access*, 7:51522–51532.

Boyang Yu, Fei Tang, Daji Ergu, Rui Zeng, Bo Ma, and Fangyao Liu. 2024. Efficient classification of malicious urls: M-bert-a modified bert variant for enhanced semantic understanding. *IEEE Access*.