DSLNLP@NLU of Devanagari Script Languages 2025: Leveraging BERT-based Architectures for Language Identification, Hate Speech Detection and Target Classification

Shraddha Chauhan
Electronics and Communication Engineering
MNNIT-Allahabad
Prayagraj, Uttar Pradesh, 211004
shraddha76830@gmail.com

Abhinav Kumar
Computer Science and Engineering
MNNIT-Allahabad
Prayagraj, Uttar Pradesh, 211004
abhik@mnnit.ac.in

Abstract

The rapid rise of social media has emphasized the spread of harmful and hateful content, making it challenging for its identification. Contextual semantics is very important as prior studies present that context level semantics is a more trustworthy indicator of hatefulness than word level semantics for detecting hate speech. This paper attempts to check the usability of transformer-based models for the identification of hate speech on code-mixed datasets, which includes Google-MuRIL, LaBSE, XLM-Roberta-base, mbert and distil-mbert. The above is largely due to its ability for high-level representations of complex and context-dense meaning. Besides this, we experiment on ensemble approach that covers all of the above models to reach out for an even higher level of performance in detection. The experiment results show the best performing macro F1-scores are reported in case of MuRIL in comparison to other implemented models.

1 Introduction

In an age in which the growth of social media is exponential for applications like X, Facebook, and ShareChat, millions of users now communicate in ways that were simply impossible in the past. It creates easily accessible, real-time conduits for information, social commentary, and open debate. At the same time, it facilitates hate speech, misinformation, and other types of offensive content (Saumya et al., 2024; Kumari and Kumar, 2023; Kumar et al., 2023, 2021b; **?**). It is unrealistic to respond to these issues alone by using manual moderation when massive volumes of data are created every second (Saumya et al., 2024, 2021). Due to this, social media outlets increasingly look towards machine learning automation-based moderation systems.

Despite the fact that tremendous progress has been made in hate speech detection with high-resource languages like English, this research work lags way behind lower-resource languages: Hindi, Marathi, Nepali, Sanskrit, and Bhojpuri. The current scenario makes it even tougher because a lot of texts on social media are found to be code-mixed, mixing languages like Hindi or Nepali with English with unique syntactic constructs. Code-mixing makes traditional NLP tasks, especially the task at hand, tougher because there is a deeper need to discern nuanced contextual cues while crossing languages so that intentions can be classified accurately (Saumya et al., 2024; Kumar et al., 2020).

The present research work focuses on three specific subtasks-one of the tasks which involve multilingual language detection between Hindi, Nepali, Marathi, Sanskrit, and Bhojpuri in the social media text. The importance of precise language identification is directly connected with the effectiveness of hate speech detection because hate speech or inflammatory messages cannot be created unless it is written in the desired language. Classify as hate speech or non-hate speech in Hindi and Nepali is the second sub-task. The third subtask is more in-depth to identify the target of the hateful statement, determining whether an individual, organization, or a community is being targeted (Thapa et al., 2025).

We used the state-of-the-art deep learning methods, focusing more on the transformer-based architecture that shows better performance for NLP-related tasks (Saumya et al., 2024; Kumar et al., 2021a). We used five transformer-based models such as mBERT, Distil-mBERT, MuRIL (Multilingual Representations for Indian Languages), LaBSE- (Language-Agnostic BERT Sentence Embedding) and XLM-RoBERTa, which explore unique linguistic diversity with each dataset, We address its tasks with the strength inherent to each model. Indeed, for example, one must notice that MuRIL is particularly strong within regionally nuanced language representations specific to Indian languages-for Hindi and Nepali texts. Its strength in generating quality, language-independent embed-

Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025), pages 289–294
January 19, 2025.

dings makes LaBSE very effective for processing multilingual and cross-lingual scenarios that are very common to our requirements.

Besides training individual models, We experimented with an ensemble strategy that combined mBERT, Distil-mBERT, and XLM-RoBERTa for classification and applied a voting mechanism among these models to make the final predictions. In this voting-based ensemble approach here, each model classifies an instance independently and derives a final prediction based on the majority voting of the three models. A label is assigned as a final prediction if two or more models agree on a classification outcome. Our approach is a contribution toward Language Identification, Hate Speech Detection and Target Identification for Hate Speech in Devanagari-script languages while also contributing to the overall advancement of the field, showing how transformer-based models are indeed effective in multilingual low-resource and code-mixed NLP tasks.

The rest of the paper is organized as follows: Section 2 lists related work, Section 3 discusses dataset & task, Section 4 discusses the proposed methodology. The outcome of the proposed model is listed in Section 5 and the paper is concluded in Section 6.

2 Related Work

Multilingual NLP has seen significant progress in recent years, with more intensive needs for effective ways and methods to handle many divergent languages, more often in regions with vast rich linguistic diversity. Detection of hate speech , target hate speech (Malik et al., 2024) or language classification has made ways through traditional statistical as well as machine learning-based methods over such techniques as Support Vector Machines SVMs, Naive Bayes classifiers, recurrent neural networks (RNNs) trained through curated language features (Liu et al., 2024; Mandal et al., 2024). These models fail to capture the contextual nuances of language and perform poorly on code-mixed and low-resource languages because of limited availability of training data and reliance on language-specific pre-processing (Conneau et al., 2019).

In our study, We used five transformer-based models individually, namely, mBERT (Devlin et al., 2018), Distil-mBERT (Sanh et al., 2019), Google's MuRIL (Khanuja et al., 2021), LaBSE and XLM-RoBERTa (Conneau et al., 2019) to check their capacity for text multilinguality processing for our study. This approach will enable the singling out of their individual strengths as well as weaknesses in either hate speech detection, target identification for hate speech or language classification tasks. In (Jafri et al., 2024a), Machine Learning algorithms like Naive Bayes, Decision Tree, SVM and Transformer based Deep Learning models like BERT, XLM-RoBERTa and Hard Ensemble of BERTs is used in original and augmented dataset of CHUNAV to analyze Hindi hate speech and targeted groups in Indian Election Discourse (Alam et al., 2024).

To facilitate our experiments and implementations, we use the ktrain framework (Maiya, 2022), which simplifies the process of developing and training deep learning models. This user-friendly library eases the inclusion of multiple architectures, such as transformer-based models, hence streamlining our research work. In general, the discussion of these state-of-the-art models (Khanduja et al., 2024) and ensemble techniques (Singhal and Bedi, 2024) greatly contributes to understanding and applying multilingual NLP, especially to challenges such as low-resource languages and code-mixed text.

3 Dataset & Task

The dataset covers multiple Devanagari-script languages and has a great variety of content features. There are numerals (e.g., 10, 2, 3), emoticons, links (e.g., https://t.co/wFKDRCF0Ny), tags like "@", English words (e.g., "Punjab Elections"), and very evocative, varied sentences in hindi, nepali, sanskrit, bhojpuri languages. The heterogeneity of this is an interesting challenge for the classification task and provides a broad basis for the testing of multilingual and mixed-content text processing models.

The Task on Natural Language Understanding of Devanagari Script Languages (Thapa et al., 2025) consists of three subtasks, which focus on critical challenges in processing languages written in the Devanagari script. These tasks are language identification, hate speech detection and the identification of targets of hate speech (Sarveswaran et al., 2025).

Class	Dataset size
Nepali	12,543
Marathi	11,034
Sanskrit	10,996
Bhojpuri	10,184
Hindi	7,660
Total	52,417

Table 1: Data distribution for subtask A

Class	Dataset size
Non-Hate Speech	16,805
Hate Speech	2,214
Total	19,019

Table 2: Data distribution for subtask B

3.1 Subtask A: Devanagari Script Language Identification

This is the subtask of language identification in text typed in the Devanagari script. The dataset includes five languages: Nepali (Thapa et al., 2023; Rauniyar et al., 2023), Marathi (Kulkarni et al., 2021), Sanskrit (Aralikatte et al., 2021), Bhojpuri (Ojha, 2019) and Hindi (Jafri et al., 2024b, 2023). There are total 52,417 train data, 11,232 validation data and 11,234 test data in Subtask A dataset. Distribution across five classes as shown in Table 1 are as Nepali (12,543), Marathi (11,034), Sanskrit (10,996), Bhojpuri (10,184) and Hindi (7,660).

3.2 Subtask B: Hate Speech Detection in Devanagari Script Language

The second subtask is to identify hate speech in sentences using the Devanagari script. Here, the dataset is annotated to indicate whether a given sentence contains hate speech. There are total 19,019 train data, 4,076 validation data and 4,076 test data in Subtask B dataset. Distribution across two classes as shown in Table 2 are as Non-Hate Speech (16,805) and Hate Speech (2,214). This annotated dataset is imbalanced and consists mainly of monolingual sentences in Nepali (Thapa et al., 2023; Rauniyar et al., 2023) and Hindi (Jafri et al., 2024b, 2023), underlining the need for proper detection mechanisms of different languages within the Devanagari script (Parihar et al., 2021).

3.3 Subtask C: Target Identification for Hate Speech in Devanagari Script

The last subtask is to identify the specific hate speech targets within individual sentences. The target categories are defined as individual, organization, or community. There are total 2,214 train data, 474 validation data and 475 test data in Subtask C dataset. There are 1,074 Individual, 856 Organization, and 284 Community text in training dataset as shown in Table 3.

Class	Dataset size
Individual	1,074
Organization	856
Community	284
Total	2214

Table 3: Data distribution for subtask C

4 Methodology

In this paper, we fine-tune five transformer-based models as shown in Figure 1 for the task of three distinct subtasks of classification on text in multiple languages. We used XLM-RoBERTa, Distil-mBERT, mBERT, LaBSE, and MuRIL for the evaluation of ability in Devanagari script languages. Each of these models was fine-tuned on the subtask datasets for 20 epochs with a batch size of 64 and a learning rate of $2 \times e^{-5}$, balancing between computation efficiency and convergence for the model.

Considering the average length of the token in the dataset was approximately 50, we set the maximum sequence length to 30 tokens with the intent of saving as much memory space without a loss of relevant information. The ktrain library utilizes preprocessing, training and evaluation of the model itself and has streamlined the entirety of the development pipeline as it has brought uniformity in handling data for the whole model.

In addition to the individual model performances, we tried to enhance its classification robustness using ensemble model as shown in Figure 2. We used mBERT, Distil-mBERT, and XLM-RoBERTa in an ensemble. The final voting was done using majority vote. This approach utilized diversified model architectures and combined relevant linguistic insights for making those predictions.

5 Results & Discussion

The Table 4 shows Accuracy (Acc), Precision (Pre), F1-score (F1) and Recall (Rec) of all five transformer models and ensemble model across subtask A, subtask B and subtask C. In evaluating the performance of models across the three sub-

Table 4: Performance metrics of various models across Subtask A, Subtask B and Subtask C

Model	Subtask A				Subtask B				Subtask C			
	Acc	Pre	F1	Rec	Acc	Pre	F1	Rec	Acc	Pre	F1	Rec
mBERT	98.80	98.71	98.69	98.67	88.39	77.54	47.74	50.39	64.63	58.65	51.07	51.68
Distil-mBERT	99.08	98.98	98.98	98.97	88.34	44.17	46.90	50.00	60.63	52.78	48.06	48.38
LaBSE	99.67	99.64	**99.64**	99.65	89.08	78.9	57.12	55.53	35.78	32.94	30.12	30.24
MuRIL	99.60	99.56	99.56	99.56	89.54	77.21	64.59	61.09	68.42	61.97	**61.01**	60.60
XLM-RoBERTa	99.53	99.46	99.48	99.50	89.57	76.49	**66.13**	62.57	66.52	59.00	58.15	57.73
Ensemble Model	99.13	99.05	99.04	99.02	88.59	75.67	51.81	52.42	69.05	63.91	57.48	56.84

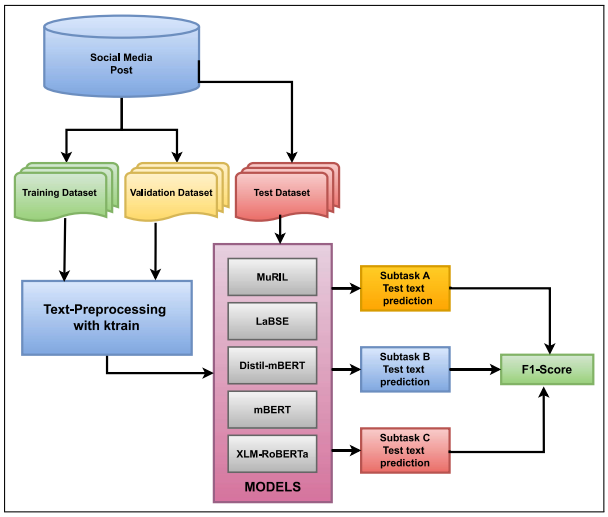

Figure 1: Flow chart of our Transformer based model.

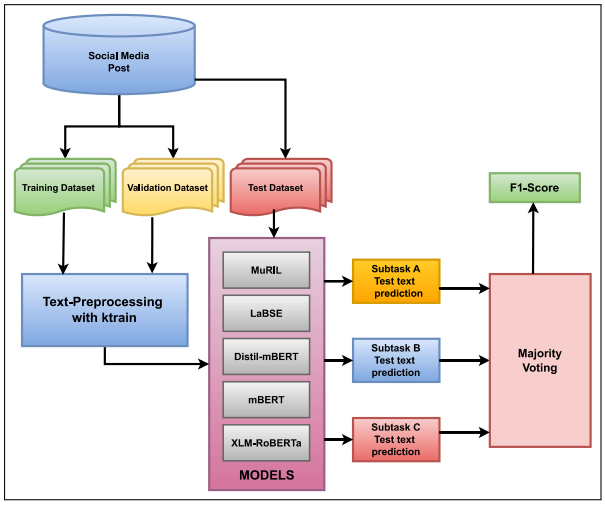

Figure 2: Flow chart of our Ensemble model.

tasks, F1 score was considered the primary comparison score, highlighting the models balanced performance in terms of precision and recall. For Subtask A that is Devanagari Script Language Identification, the LaBSE model gave better performance with an F1 score of 99.64, which showed its well-capable handling of multilingual text in the Devanagari script. In Subtask B (Hate Speech Detection), XLM-RoBERTa had the highest F1 score at 66.13, which indicated the model's ability to detect hate speech across Hindi and Nepali, where language nuances are complex. For Subtask C (Target Identification for Hate Speech), MuRIL was able to outperform other models with an F1 score of 61.01, meaning it can clearly identify hate speech targets as "individual," "organization," and "community." Although the ensemble model generated stable outcomes on subtasks, it failed to surpass individual models such as LaBSE, XLM-RoBERTa and MuRIL for tasks specific F1 scores, meaning even though ensembling makes prediction stable across task space, it is instead likely that strengths of various individual models can be aptly put to use once a proper task-specific ensemble model has been chosen rather than just going for an agnostic ensemble. The variation across tasks indicates that feature extraction and linguistic knowledge are to be used distinctly for effective results in each subtask. This is an essential insight for future work involving Devanagari script language processing and multilingual tasks in general.

6 Conclusion

The results show that transformer-based models have impressive capabilities for various NLP tasks

in Devanagari-scripted languages. Each model brought in different strengths: LaBSE was highly effective in language identification across closely related languages, XLM-RoBERTa excelled in hate speech detection, as it is cross-lingually designed; and MuRIL achieved the highest accuracy in hate speech target identification as it is pre-trained on Indian languages. Although stable result was presented by the ensemble over all the tasks, every individual model performed better at task-specific metrics than the combination. These results indicate that instead of generalized ensembling for subtler multi-linguistic applications of NLP, there is more possibility of targeting the application according to which the most competent model selection would be advantageous. For the current domain, in which this approach has given direction, similar future directions can be highlighted for researching Devanagari language processing based more on specificity of the task rather than general application techniques.

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, St. Julian's, Malta. Association for Computational Linguistics.

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Farhan Ahmad Jafri, Kritesh Rauniyar, Surendrabikram Thapa, Mohammad Aman Siddiqui, Matloob Khushi, and Usman Naseem. 2024a. Chunav: Analyzing hindi hate speech and targeted groups in indian election discourse. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

Farhan Ahmad Jafri, Kritesh Rauniyar, Surendrabikram Thapa, Mohammad Aman Siddiqui, Matloob Khushi, and Usman Naseem. 2024b. 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.

Namit Khanduja, Nishant Kumar, and Arun Chauhan. 2024. Telugu language hate speech detection using deep learning transformer models: Corpus generation and evaluation. *Systems and Soft Computing*, 6:200112.

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. *Preprint*, arXiv:2103.10730.

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.

Abhinav Kumar, Jyoti Kumari, and Jiesth Pradhan. 2023. Explainable deep learning for mental health detection from english and arabic social media posts. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Abhinav Kumar, Pradeep Kumar Roy, and Sunil Saumya. 2021a. An ensemble approach for hate and offensive language identification in english and indo-aryan languages. In *FIRE (Working Notes)*, pages 439–445.

Abhinav Kumar, Sunil Saumya, and Jyoti Prakash Singh. 2020. NITP-AI-NLP@ HASOC-Dravidian-CodeMix-FIRE2020: A machine learning approach to identify offensive languages from dravidian code-mixed text. In *FIRE (Working notes)*, pages 384–390.

Gunjan Kumar, Jyoti Prakash Singh, and Abhinav Kumar. 2021b. A deep multi-modal neural network for the identification of hate speech from social media. In *Responsible AI and Analytics for an Ethical and Inclusive Digitized Society: 20th IFIP WG 6.11 Conference on e-Business, e-Services and e-Society, I3E 2021, Galway, Ireland, September 1–3, 2021, Proceedings 20*, pages 670–680. Springer.

Jyoti Kumari and Abhinav Kumar. 2023. JA-NLP@ LT-EDI-2023: empowering mental health assessment: A roberta-based approach for depression detection. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 89–96.

Dengyi Liu, Minghao Wang, and Andrew G. Catlin. 2024. Detecting anti-semitic hate speech using transformer-based large language models. *Preprint*, arXiv:2405.03794.

Arun S. Maiya. 2022. ktrain: A low-code library for augmented machine learning. *Preprint*, arXiv:2004.10703.

Muhammad Shahid Iqbal Malik, Aftab Nawaz, and Mona Mamdouh Jamjoom. 2024. Hate speech and target community detection in nastaliq urdu using transfer learning techniques. *IEEE Access*, 12:116875–116890.

Atanu Mandal, Gargi Roy, Amit Barman, Indranil Dutta, and Sudip Kumar Naskar. 2024. Attentive fusion: A transformer-based approach to multimodal hate speech detection. *Preprint*, arXiv:2401.10653.

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

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.

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*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

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)*.

Sunil Saumya, Abhinav Kumar, and Jyoti Prakash Singh. 2021. Offensive language identification in dravidian code mixed social media text. In *Proceedings of the first workshop on speech and language technologies for Dravidian languages*, pages 36–45.

Sunil Saumya, Abhinav Kumar, and Jyoti Prakash Singh. 2024. Filtering offensive language from multilingual social media contents: A deep learning approach. *Engineering Applications of Artificial Intelligence*, 133:108159.

Kriti Singhal and Jatin Bedi. 2024. Transformers@LT-EDI-EACL2024: Caste and migration hate speech detection in Tamil using ensembling on transformers. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 249–253, St. Julian's, Malta. Association for Computational Linguistics.

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.