byteSizedLLM@NLU of Devanagari Script Languages 2025: Hate Speech Detection and Target Identification Using Customized Attention BiLSTM and XLM-RoBERTa Base Embeddings

Rohith Gowtham Kodali
ASRlytics
Hyderabad, India
rohitkodali@gmail.com

Durga Prasad Manukonda
ASRlytics
Hyderabad, India
mdp0999@gmail.com

Daniel Iglesias
Digi Sapiens
Frankfurt, Germany
diglesias@web.de

Abstract

This paper presents a novel approach to hate speech detection and target identification across Devanagari-script languages, with a focus on Hindi and Nepali. Leveraging an Attention BiLSTM-XLM-RoBERTa architecture, our model effectively captures language-specific features and sequential dependencies crucial for multilingual natural language understanding (NLU). In Task B (Hate Speech Detection), our model achieved a Macro F1 score of 0.7481, demonstrating its robustness in identifying hateful content across linguistic variations. For Task C (Target Identification), it reached a Macro F1 score of 0.6715, highlighting its ability to classify targets into "individual," "organization," and "community" with high accuracy. Our work addresses the gap in Devanagari-scripted multilingual hate speech analysis and sets a benchmark for future research in low-resource language contexts.

1 Introduction

The rapid growth of online platforms has heightened concerns around the detection and mitigation of hate speech. In the context of South Asia, where languages such as Nepali and Hindi predominantly use the Devanagari script, there is a pressing need for specialized natural language understanding (NLU) approaches that can handle the complex, multilingual nature of online discourse. Addressing these concerns, the "Shared Task on Natural Language Understanding of Devanagari Script Languages" at CHIPSAL@COLING 2025 presents a series of challenges focused on hate speech processing, specifically hate speech detection and target identification(Thapa et al., 2025; Sarveswaran et al., 2025).

Subtask B, *Hate Speech Detection in Devanagari Script Languages*, tackles the task of binary classification, aiming to identify whether a given sentence contains hate speech. The multilingual dataset, containing texts in Nepali and Hindi, underscores the need for models that can handle the nuances of each language while using a common script. This task emphasizes language-specific considerations essential for accurate detection, as hate speech often exhibits linguistic subtleties, cultural references, and slang unique to each language.

Expanding upon the hate speech detection task, Subtask C, *Target Identification for Hate Speech in Devanagari Script Languages*, introduces the challenge of identifying specific targets of hate speech. Given a hateful sentence, the task requires participants to classify the target as an *individual*, *organization*, or *community*. Target identification is crucial to understanding the nature and intended focus of hate speech, providing valuable insights that facilitate more effective responses and moderation strategies.

Our hybrid model integrates an attention-driven BiLSTM with XLM-RoBERTa embeddings to tackle hate speech detection and target identification. The attention mechanism enhances the BiLSTM's ability to focus on critical contextual cues, while XLM-RoBERTa provides robust multilingual embeddings. Together, these components enable our architecture to achieve exceptional precision, contributing to sophisticated multilingual NLU systems and fostering safer online interactions, particularly for Devanagari-scripted languages.

2 Related Work

The rise of hate speech on digital platforms has spurred research efforts in detection, yet studies for Devanagari-script languages like Hindi, Nepali, and Marathi remain limited due to script complexity and dialect diversity. Detecting hate speech in these languages is essential for fostering safer online environments. To date, research has primarily focused on monolingual hate speech detection in Hindi, Nepali, and Marathi, with some stud-

Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025), pages 242–247
January 19, 2025. ©2025 International Committee on Computational Linguistics

ies exploring multilingual hate speech detection in Hindi-Marathi and Hindi-English combinations using traditional machine learning and Transformer-based deep learning approaches. (Velankar et al., 2021; Kumari et al., 2024; Sreelakshmi et al., 2020; Sharma et al., 2022; Niraula et al., 2021; B. et al., 2019; Shukla et al., 2022; Velankar et al., 2021; T.Y.S.S and Aravind, 2019; Chavan et al., 2022; Mathur et al., 2018). However, there is a notable gap in multilingual hate speech detection specifically between Nepali and Hindi, where the combined detection remains unaddressed. Additionally, no study to date has incorporated a multilingual Devanagari-script dataset that includes target identification, categorizing targets into "individual," "organization," or "community."

3 Dataset and Task

Subtask B involves identifying whether a sentence contains hate speech in Devanagari-scripted languages, specifically Nepali(Thapa et al., 2023; Rauniyar et al., 2023) and Hindi (Jafri et al., 2024, 2023). The dataset is divided into Non-Hate and Hate categories, as shown in Table 1, requiring models to effectively detect hate speech within these languages.

Class	Train	Valid	Test
Non-Hate (0)	16805	3602	3601
Hate (1)	2214	474	745
Total	19019	4076	4076

Table 1: Distribution of samples in Train, Validation, and Test datasets for Subtask B

In Subtask C, the objective is to identify the target of hate speech, categorizing it as "individual," "organization," or "community." This task is crucial for understanding the specific direction of hate speech within the Devanagari script context. Table 2 displays the dataset distribution for each target category.

Class	Train	Valid	Test
Individual (0)	1074	230	230
Organization (1)	856	183	184
Community (2)	284	61	61
Total	2214	474	475

Table 2: Distribution of samples in Train, Validation, and Test datasets for Subtask C

Additionally, we curated datasets to fine-tune the multilingual RoBERTa model (xlm-roberta-base) for masked language modeling across five languages, following the methodology of Joshi (2022). The datasets included Bhojpuri (9.3MB from GitHub[1]), Nepali (50MB from Kaggle[2]), Sanskrit (50MB from Kaggle[3]), and both Hindi and Marathi (50MB each from AI4Bharat[4]).

4 Methodology

This study presents a hybrid Attention BiLSTM-XLM-RoBERTa model, inspired by Hochreiter and Schmidhuber (1997); Conneau et al. (2019); Manukonda and Kodali (2024a); Kodali and Manukonda (2024); Manukonda and Kodali (2024b); Brauwers and Frasincar (2023), for hate speech detection and target identification in Devanagari script. As illustrated in Figure 1, the model combines deep contextual embeddings from the fine-tuned masked language model (MLM) of XLM-RoBERTa with a BiLSTM and attention mechanism to enhance language-specific feature extraction.

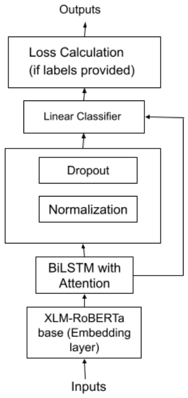

Figure 1: Architecture of the BiLSTM-XLM-RoBERTa Classifier Model. Residual components like layer normalization and dropout regularization enhance generalization.

The input sequence is first passed to XLM-RoBERTa base, generating embeddings $\mathbf{X} \in R^{T \times D}$, where $D = 768$:

$$\mathbf{X} = \mathbf{XLMRoBERTa}(input\_ids, attention\_mask) \quad (1)$$

[1] https://github.com/shashwatup9k/bho-resources

[2] https://www.kaggle.com/datasets/lotusacharya/nepalinewsdataset

[3] https://www.kaggle.com/datasets/rushikeshdarge/sanskrit

[4] https://github.com/AI4Bharat/indicnlp_corpus

These embeddings are fed into a BiLSTM, which produces bidirectional hidden states $\mathbf{H}fwd$ and $\mathbf{H}bwd$, combined as:

$$\mathbf{H}_t = [\mathbf{H}_{fwd,t}; \mathbf{H}_{bwd,t}] \tag{2}$$

An attention mechanism assigns relevance to each $\mathbf{H}_t$, generating attention weights α_t:

$$\mathbf{a}_t = \tanh(\mathbf{W}_{att} \cdot \mathbf{H}_t), \quad \alpha_t = \frac{\exp(\mathbf{a}_t)}{\sum_{t=1}^{T} \exp(\mathbf{a}_t)} \tag{3}$$

The attention-weighted representation $\mathbf{H}_{attended}$ is:

$$\mathbf{H}_{attended} = \sum_{t=1}^{T} \alpha_t \cdot \mathbf{H}_t \tag{4}$$

Layer normalization and dropout are optional residuals that mitigate overfitting and stabilize training, especially in complex language scenarios. They are applied to $\mathbf{H}_{attended}$ to enhance stability, particularly for smaller datasets:

$$\mathbf{H}_{dropout} = Dropout(LayerNorm(\mathbf{H}_{attended})) \tag{5}$$

Finally, $\mathbf{H}_{dropout}$ is passed through a classification layer to produce logits:

$$\textbf{logits} = \mathbf{W}_{cls} \cdot \mathbf{H}_{dropout} + \mathbf{b}_{cls} \tag{6}$$

The model is trained using cross-entropy loss L:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{7}$$

This architecture leverages XLM-RoBERTa embeddings, BiLSTM processing, and attention for accurate language differentiation in Devanagari-scripted contexts.

5 Experiment Setup

Our experimental setup involved data preprocessing, model fine-tuning, and architecture optimization to evaluate hate speech detection (Task B) and target identification (Task C) across Devanagari-scripted languages. Performance was assessed using accuracy and Macro F1 scores on the validation dataset.

Data preprocessing included tokenization and normalization to ensure compatibility with XLM-RoBERTa, with all text standardized to the Devanagari script. Fine-tuning on masked language modeling (MLM) used a 15% masking ratio, achieving a perplexity score of 5.33 over 7 epochs, indicating effective contextual adaptation to Devanagari-scripted languages.

After testing various classifiers, we selected an Attention BiLSTM-XLM-RoBERTa architecture (Figure 1) due to its superior performance. This model integrates XLM-RoBERTa base embeddings with a BiLSTM layer (hidden size of 256, 2 LSTM layers, and dropout rate of 0.3) to capture sequential dependencies, with an attention mechanism to emphasize language-specific and contextually relevant features. For Task B (hate speech detection), we used a learning rate of 1×10^{-5}, while for Task C (target identification), a higher learning rate of 2×10^{-5} was applied. Optional residual layers (layer normalization and dropout) were added to improve stability and mitigate overfitting.

This setup provides a robust framework for evaluating the effects of model fine-tuning, architecture, and data preparation on multilingual hate speech detection and target identification within the Devanagari script.

6 Results and Discussion

During data processing, URLs and user IDs were removed, while tweet tags were retained, as removing the tags slightly reduced F1 scores. Table 3 shows the performance of various classifiers using fine-tuned XLM-RoBERTa base embeddings on Task B and Task C. The **Attention BiLSTM-XLM-RoBERTa** model consistently outperformed other classifiers, achieving the highest F1-Scores of 0.7481 for Task B and 0.6715 for Task C. This result underscores the effectiveness of combining BiLSTM with XLM-RoBERTa base to capture sequential and contextual information essential for Devanagari-scripted language tasks. The BiLSTM with XLM-RoBERTa base embedding(BiLSTM-XLM-RoBERTa) model alone showed the F1-Scores of 0.7065 (Task B) and 0.6356 (Task C), outperforming XLM-RoBERTa base model scores of 0.6912 (Task B) and 0.6147 (Task C), demonstrating the benefits of sequential processing.

Among traditional classifiers, Logistic Regression and XGBoost delivered moderate results, with F1-Scores of 0.6528 and 0.6034 on Task B. Ensemble methods did not outperform transformer-based models, and SVC and Extra Trees showed the lowest F1-Scores, indicating limited effectiveness in handling this language data.

Our team, **byteSizedLLM**, secured 7th place

Classifier	**Task B F1-Score**	**Task C F1-Score**
XLM-RoBERTa base (Transformers)	0.6912	0.6147
XGBoost (xgb)	0.6034	0.4856
Random Forest (rf)	0.5038	0.4310
Logistic Regression (lr)	0.6528	0.5059
Gradient Boosting (gb)	0.5455	0.4760
Support Vector Classifier (svc)	0.4691	0.4171
AdaBoost (ada)	0.5684	0.4056
Extra Trees (extra_trees)	0.4896	0.4089
Ridge Classifier (ridge)	0.5626	0.4714
Stochastic Gradient Descent (sgd)	0.5813	0.4509
Ensemble (xgb, lr, rf, svc, sgd)	0.5572	0.4641
BiLSTM-XLM-RoBERTa	0.7065	0.6356
Attention BiLSTM-XLM-RoBERTa	**0.7481**	**0.6715**

Table 3: Comparison of Classifiers on Task B and Task C Test Sets

in both Task B and Task C based on F1 Macro scores, closely matching the top-ranked scores and underscoring our model's competitiveness. This strong performance highlights our approach's effectiveness, though limited open-source datasets for fine-tuning XLM-RoBERTa base on Devanagari-scripted languages like Nepali and Marathi constrained our ability to capture nuanced linguistic patterns and regional variations fully.

Fine-tuning XLM-RoBERTa's masked language model (MLM) on task-specific data significantly boosted performance, illustrating the value of tailored fine-tuning for Devanagari-scripted languages. The Attention BiLSTM-XLM-RoBERTa model successfully captured complex linguistic features by integrating attention mechanisms with BiLSTM and XLM-RoBERTa embeddings. While data limitations posed challenges, fine-tuning proved essential for adapting the model to low-resource contexts. Future research could explore larger models and expanded datasets to further improve adaptability and robustness across diverse linguistic features.

7 Conclusion

This study introduced a hybrid Attention BiLSTM-XLM-RoBERTa model for language identification in Devanagari-scripted languages, effectively integrating XLM-RoBERTa base embeddings with BiLSTM and attention mechanisms to capture both contextual and sequential features. The model's competitive F1 scores in both Task B and Task C validate this approach's effectiveness for nuanced language classification, achieving a strong 7th-place ranking in both tasks despite limited fine-tuning data.

Our findings underscore the strength of combining transformer-based embeddings with BiLSTM and attention for accurate multilingual language identification, particularly in low-resource contexts. Future work could explore larger model variants and expanded datasets to further improve performance in these settings, enhancing the model's adaptability and effectiveness across diverse linguistic features.

8 Limitations and Ethical Considerations

8.1 Limitations

The Attention BiLSTM-XLM-RoBERTa model showed promising performance, though it has limitations in generalizability. Using the XLM-RoBERTa base may limit its ability to capture complex linguistic nuances, and computational constraints restricted exploration of larger XLM-RoBERTa variants. Additionally, limited data for fine-tuning the masked language model (MLM) could impact robustness, particularly for less-represented Devanagari-scripted languages.

8.2 Ethical Considerations

This study prioritizes inclusivity for low-resource Devanagari-scripted languages, recognizing the potential impacts on linguistic communities. To address concerns of bias and fairness, we conduct regular evaluations of training data and model outputs, promote responsible interpretation and implementation of model outputs, and carefully consider community impact. These measures aim to foster fair and inclusive language technologies.

References

Premjith B., Soman Kp, and K. Sreelakshmi. 2019. Amrita cen at hasoc 2019: Hate speech detection in roman and devanagiri scripted text.

Gianni Brauwers and Flavius Frasincar. 2023. A general survey on attention mechanisms in deep learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3279–3298.

Tanmay Chavan, Shantanu Patankar, Aditya Kane, Omkar Gokhale, and Raviraj Joshi. 2022. A twitter bert approach for offensive language detection in marathi. *Preprint*, arXiv:2212.10039.

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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

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.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Rohith Kodali and Durga Manukonda. 2024. byteSizedLLM@DravidianLangTech 2024: Fake news detection in Dravidian languages - unleashing the power of custom subword tokenization with Subword2Vec and BiLSTM. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 79–84, St. Julian's, Malta. Association for Computational Linguistics.

Gitanjali Kumari, Dibyanayan Bandyopadhyay, Asif Ekbal, and Vinutha B. NarayanaMurthy. 2024. CM-off-meme: Code-mixed Hindi-English offensive meme detection with multi-task learning by leveraging contextual knowledge. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3380–3393, Torino, Italia. ELRA and ICCL.

Durga Manukonda and Rohith Kodali. 2024a. byteLLM@LT-EDI-2024: Homophobia/transphobia detection in social media comments - custom subword tokenization with Subword2Vec and BiLSTM. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 157–163, St. Julian's, Malta. Association for Computational Linguistics.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2024b. Enhancing multilingual natural language processing with custom subword tokenization: Subword2vec and bilstm integration for lightweight and streamlined approaches. In *2024 6th International Conference on Natural Language Processing (ICNLP)*, pages 366–371.

Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018. Detecting offensive tweets in Hindi-English code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26, Melbourne, Australia. Association for Computational Linguistics.

Nobal B. Niraula, Saurab Dulal, and Diwa Koirala. 2021. Offensive language detection in Nepali social media. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 67–75, Online. Association for Computational Linguistics.

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)*.

Arushi Sharma, Anubha Kabra, and Minni Jain. 2022. Ceasing hate with moh: Hate speech detection in hindi–english code-switched language. *Information Processing Management*, 59(1):102760.

Shubham Shukla, Sushama Nagpal, and Sangeeta Sabharwal. 2022. Hate speech detection in hindi language using bert and convolution neural network. In *2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pages 642–647.

K Sreelakshmi, B Premjith, and K.P. Soman. 2020. Detection of hate speech text in hindi-english code-mixed data. *Procedia Computer Science*, 171:737–744. Third International Conference on Computing and Network Communications (CoCoNet'19).

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.

Santosh T.Y.S.S and K. V. S. Aravind. 2019. Hate speech detection in hindi-english code-mixed social media text. *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*.

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