Natural Language Understanding of Devanagari Script Languages: Language Identification, Hate Speech and its Target Detection

Surendrabikram Thapa[1], Kritesh Rauniyar[2], Farhan Ahmad Jafri[3], Surabhi Adhikari[4], Kengatharaiyer Sarveswaran[5], Bal Krishna Bal[6], Hariram Veeramani[7], Usman Naseem[8]
[1]Virginia Tech, USA, [2]Delhi Technological University, India, [3]Jamia Millia Islamia, India,
[4]Columbia University, USA, [5]University of Jaffna, Sri Lanka, [6] Kathmandu University, Nepal,
[7]UCLA, USA, [8]Macquarie University, Australia
[1]surendrabikram@vt.edu, [2]rauniyark11@gmail.com,
[5]sarves@univ.jfn.ac.lk, [6]bal@ku.edu.np

Abstract

The growing use of Devanagari-script languages such as Hindi, Nepali, Marathi, Sanskrit, and Bhojpuri in digital form including social media presents unique challenges for natural language understanding (NLU), particularly in language identification, hate speech detection, and target classification. To address these challenges, we organized a shared task with three subtasks: (i) identifying the language of Devanagari-script text, (ii) detecting hate speech, and (iii) classifying hate speech targets into individual, community, or organization. A curated dataset combining multiple corpora was provided, with splits for training, evaluation, and testing. The task attracted 113 participants, with 32 teams submitting models evaluated on accuracy, precision, recall, and macro F1-score. Participants applied innovative methods, including large language models, transformer models, and multilingual embeddings, to tackle the linguistic complexities of Devanagari-script languages. This paper summarizes the shared task, datasets, and results, and aims to contribute to advancing NLU for low-resource languages and fostering inclusive, culturally aware natural language processing (NLP) solutions.

1 Introduction

Languages written in the Devanagari script, such as Hindi, Nepali, Marathi, Sanskrit, and Bhojpuri, are integral to the cultural and linguistic heritage of millions of people across South Asia and beyond. As digital technologies continue to evolve, these languages are increasingly represented in various online domains, including social media, government communications, education platforms, and digital archives. This growing digital presence reflects the linguistic diversity and cultural richness of their speakers but also presents unique challenges for natural language understanding (NLU). The development of robust computational tools for Devanagari-script languages is essential for ensuring inclusivity in global digital ecosystems (Patil et al., 2024).

Understanding and processing these languages computationally is challenging due to their grammatical and syntactic complexities, the prevalence of dialectal variations, and frequent code-switching with other languages (Gupta and Arora, 2022). While there has been substantial progress in NLU for high-resource languages like English, many low-resource languages, including those in Devanagari script, remain underexplored (Rauniyar et al., 2023). This gap is exacerbated by the scarcity of high-quality annotated datasets and the limited adaptability of existing models designed primarily for English or other high-resource languages.

Hate speech detection and language identification are two critical NLU tasks for Devanagari-script languages. Beyond their application in social media moderation, these tasks are vital for promoting inclusive communication, safeguarding digital platforms from harmful content, and supporting broader societal goals such as equitable access to technology. Accurate language identification serves as the foundation for effective language-specific interventions, while understanding hate speech and its targets ensures the development of safer, more culturally sensitive tools for digital spaces.

To address these challenges, we organized a shared task that focuses on three critical NLU tasks for Devanagari-script languages: (i) language identification, (ii) hate speech detection, and (iii) hate speech target classification. Subtask A aims to identify the language of a given text written in Devanagari script among Nepali, Marathi, Sanskrit, Bhojpuri, and Hindi. Subtask B focuses on detecting whether a given text contains hate speech, addressing the growing need to combat online toxicity. Subtask C delves deeper, seeking to classify the target of hate speech into predefined

Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025), pages 71–82
January 19, 2025.

categories, such as individual, community, or organization.

This shared task fosters advancements in low-resource NLP research, encouraging the development of models that are not only linguistically robust but also culturally aware. In this report, we provide an overview of the shared task, detailing its structure, datasets, and evaluation metrics. We also summarize the methodologies employed by participants, the results they achieved, and the lessons learned from this initiative. Through this effort, we aim to contribute to a broader understanding of multilingual NLP and support the creation of inclusive and equitable digital technologies for underrepresented languages.

2 Shared Task Description

In this shared task, we focus on exploring the capabilities of language models and classification systems to address three distinct challenges related to Devanagari script languages. The goal is to promote advancements in NLU for low-resource languages, which are often underrepresented in mainstream NLP research.

The shared task is divided into three subtasks, each aimed at tackling a specific aspect of language processing within the Devanagari script. Participants are encouraged to develop robust and generalizable models that can handle variations in dialects, mixed-language content, and context-specific nuances that are common in social media texts written in Devanagari script. Further details on subtasks can be found below:

2.1 Subtask A: Devanagari Script Language Identification

This subtask involves determining whether a given text is in Devanagari script or not. The dataset consists of text that has been annotated to determine the language it belongs to among Nepali, Marathi, Sanskrit, Bhojpuri, and Hindi. This task focuses on accurate language recognition in multilingual contexts.

2.2 Subtask B: Hate Speech Detection in Devanagari Script Language

The purpose of this subtask is to determine whether a given text in the Devanagari script contains hate speech. The dataset comprises single utterances in Hindi and Nepali that have been marked as either containing hate speech or not. The dataset is further precisely divided into two classes: texts that have been categorized as hate speech and texts that have been categorized as non-hate speech.

2.3 Subtask C: Target Identification for Hate Speech in Devanagari Script Language

This subtask aims to identify the target audience of hate speech within a specified set of hateful text in Devanagari script. Classifying three specific targets listed in the dataset is the explicit focus of the subtask, despite the fact that hate speech texts may contain various potential targets across several categories. The texts in the dataset are labeled according to their targets, which can be classified as community, individual, or organization. Therefore, our goal is to identify these particular targets in Devanagari texts that contain hate speech. Understanding the precise nature and direction of hate speech requires completing this subtask.

3 Dataset

We conducted a total of three subtasks. Subtask A focused on identifying five different Devanagari languages and utilized six datasets. For Nepali, we used two datasets: NEHATE (Thapa et al., 2023) and NAET (Rauniyar et al., 2023). The Marathi language was represented by the L3CubeMahaSent dataset (Kulkarni et al., 2021), and the Sanskrit language by the Itihasa dataset (Aralikatte et al., 2021). Additionally, we used a dedicated Bhojpuri dataset (Ojha, 2019), and for Hindi, we employed the IEHate dataset (Jafri et al., 2023). A total of 52,422 rows of data were used for the training set, 11,233 rows for the evaluation set, and 11,234 rows for the test set. Subtask B focused on hate speech detection and utilized three datasets: NEHATE, NAET, and IEHate. Additionally, Subtask C, which aims to identify targets of hate speech, also employed the NEHATE and NAET datasets. For Subtask C, we further included the CHUNAV dataset (Jafri et al., 2024). For each subtask, we stratified the dataset into stages for training, evaluation, and testing, maintaining a proportional split ratio of around 70-15-15. Table 1 represents the dataset statistics for the shared task.

4 Participants' Methods

In this section, we describe the various methods used by the participants who submitted the system

Subtask	Classes	Train	Eval	Test	Total
Subtask A	Nepali	12,544	2,688	2,688	17,920
	Marathi	11,034	2,364	2,365	15,763
	Sanskrit	10,996	2,356	2,356	15,708
	Bhojpuri	10,184	2,182	2,183	14,549
	Hindi	7,664	1,643	1,642	10,949
Subtask B	Hate	2,214	474	475	3,163
	Non-Hate	16,805	3,602	3,601	24,008
Subtask C	Individual	1,074	230	230	1,534
	Organization	856	183	184	1,223
	Community	284	61	61	406

Table 1: Dataset statistics for our shared task.

description paper.

4.1 Overview

Out of the 113 participants who registered for the shared task, a total of 25 participants submitted scores for subtask A, 32 participants for subtask B, and 27 participants for subtask C. The leaderboards for these subtasks are provided in Table 2, Table 3, and Table 4. In subtask A, team CUFE (Ibrahim, 2025) achieved the highest performance with an impressive F1-score of 99.97. Similarly, in subtask B, Paramananda (Acharya et al., 2025) secured the top position with an F1-score of 91.36, while in subtask C, MDSBots (Thapaliya et al., 2025) emerged as the leader with the highest F1-score of 76.84.

4.2 Methods

Below, we provide a summary of the system descriptions provided by the participating teams in the shared task. These summaries are derived from the approaches detailed by the participants in their system description papers.

4.2.1 Subtask A

CUFE (Ibrahim, 2025) utilized fastText classifier for language identification, leveraging its subword modeling capabilities through n-grams along with systematic token generation using the tokenizer by Team et al. (2022). The proposed system achieves a near-perfect F1 score of 99.97% on the test set and secures the first position in the shared task.

1-800-SHARED-TASKS (Purbey et al., 2025) utilized ensemble model with IndicBERT V2 (Doddapaneni et al., 2022) and achieved an exceptional F1-score of 99.79% and secured third position on leaderboard. Individual models like MuRIL (Khanuja et al., 2021) and Gemma-2 (Team et al., 2024) also delivered strong performances. The results demonstrate the effectiveness of multilingual transformer models in distinguishing between Devanagari-script languages. Ensembling enhanced robustness and reduced misclassifications, leveraging complementary strengths of individual models to achieve near-perfect classification accuracy.

byteSizedLLM (Manukonda and Kodali, 2025) used a hybrid Attention BiLSTM-XLM-RoBERTa model that achieved an F1-score of 99.74%.

MDSBots Thapaliya et al. (2025), used transformer models and TF-IDF feature extractor methods in conjunction with traditional machine learning models to achieve optimal outcomes. They finetuned the mBERT, XLM-R-Base, XLM-RoBERTa-Large, Varta-BERT, MuRIL-Base, and MURTweet transformer models. They applied the undersampling strategy, in which models were trained on half of the task's total data, to address the issue of class imbalance. Using MURTweet, they were able to attain a maximum f1-score and recall of 99.68% and precision of 99.67%. Out of all the competing teams, they achieved the sixth-best ranking on this subtask-A.

Anisan (Shanto et al., 2025) used an ensemble method that leverages the strengths of multiple transformers namely mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020) and IndicBERT (Doddapaneni et al., 2023) to achieve 99.68% accuracy.

DSLNLP (Chauhan and Kumar, 2025) employed mBERT, Distil-mBERT, and XLM-RoBERTa models in an ensemble approach to attain a higher F1-score; however, LaBSE provides the highest performance on this task. To achieve the best results, they first optimized the bert variants, XLM-RoBERTa, DistilmBERT, mBERT, LaBSE, and MuRIL, on the Devanagari script dataset. To make the predictions, they included important linguistic insights and employed a variety of model designs using the majority vote in the ensembling approach. On LaBSE, they achieve the highest f1-score, recall, and precision of 99.64%, 99.65%, and 99.64%, respectively. Their LaBSE model placed eighth out of all the teams who took part in subtask A.

Rank	Team Name	Codalab Username	Accuracy (%)	Recall (%)	Precision (%)	F1-score (%)
1	CUFE (Ibrahim, 2025)	michaelibrahim	99.97	99.97	99.97	99.97
2	CLTL	Yestin	99.82	99.82	99.82	99.84
3	1-800-SHARED-TASKS (Purbey et al., 2025)	jebish7	99.79	99.81	99.80	99.82
4	1-800-SHARED-TASKS (Purbey et al., 2025)	lazyboy.blk	99.76	99.76	99.76	99.79
5	byteSizedLLM (Manukonda and Kodali, 2025)	mdp0999	99.75	99.73	99.74	99.76
6	MDS Bots (Thapaliya et al., 2025)	sumanpaudel	99.68	99.67	99.68	99.72
7	AniSan (Shanto et al., 2025)	Priya57	99.66	99.64	99.65	99.69
8	DSLNLP (Chauhan and Kumar, 2025)	Abhinav05	99.65	99.64	99.65	99.68
9	-	sandeep_S	99.58	99.59	99.58	99.63
10	Nepali Transformers (Khadka et al., 2025)	Pilot-Khadka	99.56	99.54	99.55	99.60
11	-	decem	99.56	99.55	99.55	99.60
12	-	jerrytomy	99.50	99.55	99.53	99.57
13	CUET_Big_O (Hossan et al., 2025)	dark_shadow	99.40	99.41	99.41	99.47
14	SKPD Emergency (Shakya et al., 2025)	shubham_shakya	99.44	99.38	99.41	99.48
15	Paramananda (Acharya et al., 2025)	sure	99.39	99.40	99.39	99.46
16	Nitro NLP	menta27	99.38	99.37	99.38	99.46
17	NLP Champs	abhay-43	99.34	99.34	99.34	99.40
18	Paramananda (Acharya et al., 2025)	fulbutte	99.18	99.18	99.18	99.26
19	-	samanjoy2	99.11	99.10	99.11	99.18
20	AGRJ	getabhi89	96.78	96.43	96.49	96.90
21	-	Tanvir_77	96.30	96.19	96.14	96.44
22	-	RohanR	95.69	95.77	95.72	95.99
23	CUET_Big_O (Hossan et al., 2025)	sakib07	94.24	95.11	94.54	95.08
24	CipherLoom	Nikhil_7280	65.72	57.19	59.71	69.87
25	CNLP-NITS	advaitha	56.70	67.72	50.46	53.53

Table 2: Sub-task A (Devanagari Script Language Identification) Leaderboard, Ranked by Macro F1-score. All scores are presented as percentages (%). It is to be noted that this leaderboard contains the score till the test deadline and does not consider further runs done by participants as a part of the system description paper.

Nepali Transformers (Khadka et al., 2025) used the Twitter-trained multilingual RoBERTa (Barbieri et al., 2020) and achieved exceptional results, with F1-score reaching 99.55%, outperforming other baseline models, including general-purpose and Devanagari-specific architectures. This approach secured tenth position on the leaderboard, demonstrating the model's effectiveness in handling the linguistic diversity and complexity of the Devanagari script.

CUET_Big_O (Hossan et al., 2025) used Traditional ML models such as Logistic Regression, SVM, Multinomial Naive Bayes, Random Forest and Deep Learning models such as CNN, LSTM, BiLSTM, along with the aggregation of models like CNN+GRU, CNN+BiLSTM. The best-performing model was CNN with BiLSTM which achieved an F1-score of 99.41%.

SKPD Emergency (Shakya et al., 2025), used an innovative approach using Continuous Bag of Words (CBOW) embeddings and an attention-enhanced Bidirectional Long Short-Term Memory (BiLSTM) neural network to identify languages written in Devanagari script. The results were impressive, with the model achieving a remarkable 99% overall accuracy. Sanskrit was perfectly classified, while some challenges remained in differentiating between highly similar languages like Hindi and Bhojpuri. The CBOW embeddings significantly outperformed character-level encoding, demonstrating their ability to capture semantic relationships and linguistic subtleties that character-based approaches miss.

Paramananda (Acharya et al., 2025) employed the FastText and BERT models, achieving exceptional performance with F1-scores of 99.17% and 99.39%, respectively, securing the seventeenth position on the leaderboard. While BERT marginally outperformed FastText by leveraging its deep contextual embeddings to capture nuanced linguistic differences, FastText demonstrated higher computational efficiency, making it more suitable for large-scale applications.

CUET_Big_O (Hossan et al., 2025) used CNN with BiLSTM to obtain F1-score of 99.41%. This however differs from the official leaderboard.

4.2.2 Subtask B

Paramananda Acharya et al. (2025) utilized FastText and demonstrated superior performance, particularly with data augmentation, achieving an F1 score of 81.39% and scoring first position on the leaderboard. This outperformed BERT, which struggled with an F1 score of 0.5763. Despite its contextual embedding strengths, BERT's underperformance was attributed to overfitting on sparse

Rank	Team Name	Codalab Username	Accuracy (%)	Recall (%)	Precision (%)	F1-score (%)
1	Paramananda (Acharya et al., 2025)	fulbutte	85.52	78.47	81.39	91.36
2	CLTL	Yestin	81.27	74.61	77.3	89.35
3	MDS Bots (Thapaliya et al., 2025)	sumanpaudel	76.94	76.32	76.63	90.26
4	1-800-SHARED-TASKS (Purbey et al., 2025)	jebish7	74.41	79.25	76.52	91.12
5	1-800-SHARED-TASKS (Purbey et al., 2025)	lazyboy.blk	73.6	78.95	75.88	90.97
6	-	MuhammadArham	72.61	78.14	74.94	90.68
7	byteSizedLLM (Rohith Gowtham Kodali and Iglesias, 2025)	mdp0999	77.45	72.86	74.81	88.57
8	Dll5143A (Yadav and Singh, 2025)	Dll5143A	75.76	73.45	74.52	88.98
8	Dll5143A (Yadav and Singh, 2025)	Dll5143	75.76	73.45	74.52	88.98
9	LLMsAgainstHate (Sidibomma et al., 2025)	rushendra910	71.19	78.84	74.19	90.75
10	-	jerrytomy	73.14	74.12	73.61	89.35
11	1-800-SHARED-TASKS (Purbey et al., 2025)	Siddartha-10	70.34	78.95	73.59	90.7
12	-	sandeep_S	74.63	72.56	73.52	88.59
13	CUET_HateShield Aodhora et al. (2025)	Sumaiya_127	73.52	72.38	72.93	88.57
14	Nepali Transformers (Khadka et al., 2025)	Pilot-Khadka	73.24	72	72.59	88.4
15	-	srikarkashyap	73.29	71.87	72.54	88.32
16	-	decem	66.5	76.64	69.89	89.89
17	NLPineers (Guragain et al., 2025)	anmol2059	77.62	66.39	69.14	82.58
18	CUET_823	ratnajit_dhar	67.28	71.6	69.07	88.52
19	NLP_Ninjaas	Nadika	68.77	69.34	69.04	87.44
20	CUFE (Ibrahim, 2025)	michaelibrahim	65.45	73.12	68.17	89.01
21	CIOL (Gupta et al., 2025)	azminewasi	65.47	71.06	67.62	88.4
22	NLP Champs	abhay-43	68.16	64.77	66.14	84.42
23	DSLNLP (Chauhan and Kumar, 2025)	Abhinav05	62.57	76.49	66.13	89.57
24	CUET_Big_O (Hossan et al., 2025)	dark_shadow	67.98	63.48	65.1	83.15
25	SKPD Emergency (Shakya et al., 2025)	shubham_shakya	62.62	64.03	63.26	85.62
26	AniSan (Shanto et al., 2025)	Priya57	59.47	70.69	62.06	88.44
27	-	RohanR	58.22	66.37	60.2	87.54
28	-	Tanvir_77	66.66	58.62	58.94	74.19
29	Paramananda (Acharya et al., 2025)	sure	55.9	73.77	57.63	88.76
30	CNLP-NITS	advaitha	50	44.17	46.91	88.35
31	Nitro NLP	menta27	51.84	50.81	46.49	60.28

Table 3: Sub-task B (Hate Speech Detection) Leaderboard, Ranked by Macro F1-score. All scores are presented as percentages (%). It is to be noted that this leaderboard contains the score till the test deadline and does not consider further runs done by participants as a part of the system description paper.

datasets, as evidenced by a higher evaluation score that did not generalize to test data.

MDSBots Thapaliya et al. (2025), experimented with identical transformer and classical models that were used in subtask-A. In order to reduce class imbalance, samples from the non-hate class were prioritized above samples from the majority class. They achieved a maximum f1-score, recall, and precision of 76.62%, 76.87%, and 76.38%, respectively, using MURTweet. Out of all the participating teams, they ranked third on this subtask-B.

1-800-SHARED-TASKS (Purbey et al., 2025) opted for the same ensemble technique as Subtask A and the ensemble of models achieved the highest F1-score of 0.7588 and placed fifth position on the leaderboard. Fine-tuning with focal loss (Lin, 2017) was instrumental in addressing class imbalance, and improving the detection of minority instances.

byteSizedLLM (Rohith Gowtham Kodali and Iglesias, 2025) focused on a hybrid Attention BiLSTM-XLM-RoBERTa architecture which utilized BiLSTM's sequential processing, attention mechanisms for contextual emphasis, and XLM-RoBERTa embeddings for multilingual adaption. The model attained an F1-score of 74.81% and secured seventh position on the leaderboard, surpassing other models.

Dll5143A (Yadav and Singh, 2025) used the Hierarchical Gated Adaptive Attention (HGAA) model, leveraging XLM-RoBERTa embeddings, achieved a competitive F1-score of 74.52% and securing an eighth position on the leaderboard. This model balanced precision and recall, demonstrating its robustness in detecting hate speech in Devanagari-scripted languages. Comparatively, the non-gated architecture showed lower performance. The introduction of gating mechanisms significantly improved the models ability to handle class imbalance, enhancing minority class detection at the cost of some false positives.

LLMsAgainstHate (Sidibomma et al., 2025) used Nemo-Instruct-2407 model (AI and NVIDIA) and achieved the highest performance with an F1-score of 74.52%, outperforming alternatives such as Phi-3-medium (Abdin et al., 2024) and Llama-3.1. Despite significant class imbalance favoring non-hate class, Nemo demonstrated robust detection capabilities, particularly benefiting from Parameter-Efficient Fine-Tuning using Low-Rank Adaptation (Hu et al., 2021).

CUETHateShield Aodhora et al. (2025) used the classical, deep learning, and transformer models to experiment with this task. In classical machine learning models, they incorporated the Logistic Regression, Support Vector Machine, and Random Forest model with TF-IDF feature extractor, and for deep learning models, they used CNN, Bi-LSTM, and CNNBiLSTM model with fastText and keras embedding. In transformer model, they used mBERT, MuRIL, IndicBERT, Indic-SBERT, and XLM-RoBERTa. To obtain higher-quality data, they eliminated noise in the preprocessing step, which included punctuation, emojis, hyperlinks, alphanumeric letters, and special symbols (such as slashes, brackets, and ampersands). They obtain an f1-score of 74% on XLM-RoBERTa, recall of 75% on Indic-SBERT, and precision of 72% on MuRIL.After competing against all teams, they came in at number eleven on this subtask.

Nepali Transformers (Khadka et al., 2025) deployed the Twitter-trained multilingual RoBERTa model (Barbieri et al., 2020), which achieved an F1-score of 72.93% and secured the fourteenth position on the leaderboard, outperforming general-purpose and Devanagari-specific models. This model excelled due to its domain-specific pretraining on social media datasets, effectively capturing nuanced hate speech patterns in Devanagari-scripted languages.

NLPineers (Guragain et al., 2025) used an ensemble of multilingual BERT to achieve a recall of 77.62% (ranked 3rd out of 31 in terms of recall and 17th out of 31 for an F1 score of 69.14%). To address the class imbalance, the authors used back-translation for data augmentation and cosine similarity to preserve label consistency after augmentation.

IITR-CIOL (Gupta et al., 2025) developed a model called *MultilingualRobertaClass*, which is a deep neural network built on the pre-trained IBM transformer model "ia-multilingual-transliterated-roberta". The model achieved an accuracy of 82.21%, a weighted precision of 79.84%, a weighted Recall of 82.21%, and a weighted F1 Score of 80.97%.

DSLNLP (Chauhan and Kumar, 2025), XLM-RoBERTa performs the best on this problem despite using the same ensembling algorithm as subtask-A. For the best results on hate speech recognition on Devanagari scripts, they refined the same models of Bert variations on all of the scripts, just like in the prior task. They achieved the highest precision of 78.9% by LaBSE, whereas the maximum f1-score and recall on XLM-RoBERTa were 66.13% and 62.57%, respectively. In subtask B, their XLM-RoBERTa model ended at number 23 out of all the teams that took part.

4.2.3 Subtask C

MDSBots (Thapaliya et al., 2025), employed the same models as the previous tasks, but they added a hybrid model in this task. In the hybrid model, they integrated named entity information (NER) into features produced by BERT models and used open-source large language models to reclassify samples with low confidence scores by prompting. They employed data augmentation to address the class disparity, by augmenting the minority class to represent community targets. Their best f1-score, recall, and precision on NERMURTweet were 70.98%, 70.38%, and 71.75%, respectively. Out of all the teams, they won first place for this subtask-C.

CUET_INSights (Tofa et al., 2025) combine traditional ML and Deep Learning Techniques while leveraging the multilingual capabilities of Indic-BERT & m-BERT with the adoption of Bhojpuri-to-Hindi translation along with class weights to mitigate imbalance. By utilizing these techniques including the almost perfect blend of deeper embeddings with shallow ML (TFIDF) features, the authors achieve a high F1 score of 69.17% thereby securing the third spot in the leaderboard.

CUET_Big_O (Hossan et al., 2025) used GridSearchCV for hyperparameter tuning and tested different kernels (linear, RBF) for SVM. They also tested multiple transformers: m-BERT, Indic-BERT, MuRIL-BERT, XLM-R, and Verta-BERT. The best performer was MuRIL-BERT with an F1-score of 68.32%.

1-800-SHARED-TASKS (Purbey et al., 2025) employed Gemma-2 27B model, fine-tuned using

Rank	Team Name	Codalab Username	Accuracy (%)	Recall (%)	Precision (%)	F1-score (%)
1	MDS Bots (Thapaliya et al., 2025)	sumanpaudel	70.38	71.75	70.98	76.84
2	-	Siddartha-10	68.73	74.11	70.33	77.89
3	CUET_INSights (Tofa et al., 2025)	Tofa	67.17	74.18	69.17	76.63
4	CUET_Big_O (Hossan et al., 2025)	sakib07	68.13	68.61	68.32	74.53
5	1-800-SHARED-TASKS (Purbey et al., 2025)	jebish7	66.69	71.83	68.04	76.63
6	One_by_zero (Chakraborty et al., 2025)	Dola_Chakraborty	68.1	67.88	67.98	73.68
7	byteSizedLLM (Rohith Gowtham Kodali and Iglesias, 2025)	mdp0999	66.89	67.44	67.15	74.11
8	-	jerrytomy	66.68	66.29	66.41	73.05
9	-	sandeep_S	65.72	67.54	66.37	73.89
10	CLTL	Yestin	65.46	67.4	66.12	74.53
11	Dll5143A (Yadav and Singh, 2025)	Dll5143A	65.37	66.38	65.76	71.37
12	-	srikarkashyap	64.78	67.58	65.69	72.42
13	Dll5143	Dll5143	64.75	64.76	64.74	72.21
14	LLMsAgainstHate (Sidibomma et al., 2025)	rushendra910	63.36	65.29	64.08	72
15	CipherLoom	Nikhil_7280	61.4	64.47	62.37	71.16
16	Nepali Transformers (Khadka et al., 2025)	Pilot-Khadka	62.36	61.57	61.83	68.63
17	DSLNLP (Chauhan and Kumar, 2025)	Abhinav05	60.61	61.98	61.01	68.42
18	-	decem	59.39	63.81	59.96	71.16
19	CUET_SSTM	aref111n	59.39	64.39	59.73	69.89
20	CIOL (Gupta et al., 2025)	azminewasi	58.39	59.1	58.16	66.11
21	Paramananda (Acharya et al., 2025)	sure	57.44	58.57	57.85	66.95
22	Paramananda (Acharya et al., 2025)	fulbutte	53.3	56.67	53.74	63.58
23	NLP Champs	abhay-43	50.77	51.16	50.57	58.32
24	CUFE	michaelibrahim	50.27	54.55	50.08	62.53
25	-	RohanR	45.77	56.41	44.22	60.42
26	-	Tanvir_77	46.85	41.1	43.74	61.68
27	AniSan (Shanto et al., 2025)	Priya57	45.33	44.94	42.07	61.68

Table 4: Sub-task C (Target Identification for Hate Speech) Leaderboard, Ranked by Macro F1-score. All scores are presented as percentages (%). It is to be noted that this leaderboard contains the score till the test deadline and does not consider further runs done by participants as a part of the system description paper.

ORPO (Hong et al., 2024), achieved the highest F1-score of 68.04%, with recall and precision scores of 66.69% and 71.83%, respectively. This model outperformed alternatives like Gemma-2 9B and XLM-RoBERTa, which scored lower due to challenges in detecting nuanced targets like “community". The results highlighted the model’s strong performance in identifying targets such as “individual" and “organization," while community-target detection remained a challenge, underscoring the need for richer datasets.

One_by_zero (Chakraborty et al., 2025) utilized traditional machine learning models such as Logistic Regression, SVM leveraging TF-IDF Feature Extraction, deep learning (DL) architectures such as CNN, BiLSTM, CNN+BiLSTM hybrid) utilizing Word2Vec and FastText embeddings and Transformer-based architectures such as IndicBERT, MuRIL, XLM-R while adopting oversampling for underrepresented classes. Specifically, along with IndicBERT they achieve a notable high score of 67.85% percentage securing the sixth spot in the leaderboard.

byteSizedLLM (Rohith Gowtham Kodali and Iglesias, 2025) opted Attention BiLSTM-XLM-RoBERTa architecture achieved a macro F1-score of 67.15% and also secured a seventh position on the leaderboard, effectively categorizing hate speech targets as individual, organization, or community. This model outperformed baseline approaches, with the BiLSTM-XLM-RoBERTa variant scoring 63.56% and the XLM-RoBERTa base scoring 61.47%. The attention mechanism improved focus on critical context, enhancing accuracy for complex multilingual tasks.

Dll5143A (Yadav and Singh, 2025) implemented the HGAA model which achieved a macro F1-score of 65.76% and eleventh position on the leaderboard, demonstrating effective classification for individual and organizational targets. Community target detection remained challenging due to nuanced language and class imbalance. The inclusion of gating mechanisms improved performance compared to simpler architectures, particularly in minority class detection, showcasing the model’s ability to balance precision and recall across diverse linguistic contexts.

LLMsAgainstHate (Sidibomma et al., 2025), the Nemo-Instruct-2407 model (AI and NVIDIA) delivered the strongest results with an F1-score of 64.08%. It outperformed models like Phi-3-medium and Qwen2.5 (Yang et al., 2024), showcasing its robustness in handling target-specific classifications. However, a detailed class-wise breakdown revealed a notable performance disparity, with high accuracy in detecting

“Individual" and “Organization" targets, but a significant drop for “Community". This discrepancy underscores the challenge posed by imbalanced datasets and under-represented categories.

For **Nepali Transformers** (Khadka et al., 2025), the Twitter-trained multilingual RoBERTa (Barbieri et al., 2020) demonstrated competitive performance, achieving an F1-score of 61.83%. While the model excelled in identifying “Individual" and “Organization" targets, it struggled with “Community" due to the scarcity of labeled examples. Augmentation strategies using multilingual embeddings provided modest improvements but did not fully resolve the imbalance challenges.

DSLNLP (Chauhan and Kumar, 2025), their MuRIL model performs the best on this task. In order to achieve the greatest results on target identification for hate speech on Devanagari scripts, they improved the same models of Bert variations and applied the same ensembling techniques as in the prior tasks. The ensemble method with a majority voting strategy yielded the highest precision of 63.91%, whereas the maximum F1-score and recall on MuRIL were 61.01% and 60.60%, respectively. Their MuRIL model came in at number seventeen out of all the teams who took part in subtask C.

IITR-CIOL (Gupta et al., 2025) built a model that was pre-trained on a multilingual transformer model to handle the linguistic diversity and complexity of South Asian languages. While the model performed exceptionally well in Subtask B (hate speech detection), achieving an accuracy of 88.40%, its performance in Subtask C was notably lower, with an accuracy of 66.11%. The ablation studies revealed that sequence length was the most critical factor in model performance, with longer sequences providing better context and more accurate predictions.

Paramananda (Acharya et al., 2025) used BERT which demonstrated superior performance with an F1 score of 53.74%. BERT’s success is attributed to its ability to leverage deep contextual embeddings, enabling the identification and differentiation of nuanced targets such as individuals, organizations, and communities.

5 Discussion

The shared task on Devanagari-script languages offered unique insights into the complexities of NLU for low-resource languages. Participants showcased a diverse range of approaches, from classical machine learning models to transformer-based architectures and large language models, each with distinct strengths and limitations. Success of models like IndicBERT and XLM-RoBERTa in Subtask A underscores importance of multilingual and domain-specific embeddings in effectively distinguishing between linguistically similar languages.

However, the results also illuminate several challenges. Despite achieving high overall accuracy, many models struggled with underrepresented classes, such as ‘Community’ in Subtask C, pointing to limitations of existing datasets and the need for better class balance and data augmentation techniques. Additionally, while transformer-based models excelled in capturing contextual nuances, their reliance on large-scale training data highlights the necessity of domain-specific pretraining and fine-tuning strategies tailored for low-resource languages. Moving forward, fostering collaboration within the NLP community and developing more comprehensive datasets will be crucial to addressing these challenges and advancing research in Devanagari-script languages.

6 Conclusion

This shared task on NLU for Devanagari-script languages addressed critical challenges in language identification, hate speech detection, and target classification. Through the participation of diverse teams and methodologies, the task highlighted the potential of transformer-based models, ensemble approaches, and hybrid architectures in tackling the linguistic and contextual intricacies of low-resource languages. Substantial progress was demonstrated, particularly in language identification, where models achieved near-perfect scores, showcasing the effectiveness of multilingual embeddings and pretraining on diverse datasets. Nevertheless, challenges such as class imbalance, underrepresentation of specific categories, and the need for domain-specific pretraining were identified as key areas requiring further research. This task has laid the groundwork for future exploration and highlighted the importance of more research in inclusive and culturally aware NLP solutions.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Darwin Acharya, Sundeep Dawadi, Shivram Saud, and Sunil Regmi. 2025. Paramananda@nlu of devanagari script languages 2025: Detection of language, hate speech and targets using fasttext and bert. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi. International Committee on Computational Linguistics (ICCL).

Mistral AI and NVIDIA. Title of webpage. https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407. Accessed: 2024.

Sumaiya Rahman Aodhora, Shawly Ahsan, and Mohammed Moshiul Hoque. 2025. Cuet_hateshield @ nlu of devanagari script languages 2025: Transformer-based hate speech detection in devanagari script languages. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi. International Committee on Computational Linguistics (ICCL).

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.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63.

Dola Chakraborty, Jawad Hossain, and Mohammed Moshiul Hoque. 2025. One_by_zero@nlu of devanagari script languages 2025: Target identification for hate speech leveraging transformer-based approach. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi. International Committee on Computational Linguistics (ICCL).

Shraddha Chauhan and Abhinav Kumar. 2025. Dslnlp@nlu of devanagari script languages 2025: Leveraging bert-based architectures for language identification, hate speech detection and target classification. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi. International Committee on Computational Linguistics (ICCL).

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2022. Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages. *arXiv preprint arXiv:2212.05409*.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Samridhi Gupta and Bhavna Arora. 2022. Stemming techniques on english language and devanagari script: A review. *Recent Innovations in Computing: Proceedings of ICRIC 2021, Volume 1*, pages 541–550.

Siddhant Gupta, Siddh Singhal, and Azmine Toushik Wasi. 2025. Iitr-ciol@nlu of devanagari script languages 2025: Multilingual hate speech detection and target identification in devanagari-scripted languages. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi. International Committee on Computational Linguistics (ICCL).

Vikram Gupta, Sumegh Roychowdhury, Mithun Das, Somnath Banerjee, Punyajoy Saha, Binny Mathew, Animesh Mukherjee, et al. 2022. Multilingual abusive comment detection at scale for indic languages. *Advances in Neural Information Processing Systems*, 35:26176–26191.

Anmol Guragain, Nadika Poudel, Rajesh Piryani, and Bishesh Khanal. 2025. Nlpineers@ nlu of devanagari script languages 2025: Hate speech detection using ensembling of bert-based models. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi. International Committee on Computational Linguistics (ICCL).

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Reference-free monolithic preference optimization with odds ratio. *arXiv e-prints*, pages arXiv–2403.

Md. Refaj Hossan, Nazmus Sakib, Md. Alam Miah, Jawad Hossain, and Mohammed Moshiul Hoque. 2025. Cuet_big_o@nlu of devanagari script languages 2025: Identifying script language and detecting hate speech using deep learning and transformer model. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi. International Committee on Computational Linguistics (ICCL).

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Michael Ibrahim. 2025. Cufe@nlu of devanagari script languages 2025: Language identification using fasttext. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi. International Committee on Computational Linguistics (ICCL).

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.

Pilot Khadka, Ankit BK, Ashish Acharya, Bikram K.C., Sandesh Shrestha, and Rabin Thapa. 2025. Nepali transformersnlu of devanagari script languages 2025: Detection of language, hate speech and targets. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi. International Committee on Computational Linguistics (ICCL).

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv 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.

T Lin. 2017. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*.

Vidula Magdum, Omkar Dhekane, Sharayu Hiwarkhedkar, Saloni Mittal, and Raviraj Joshi. 2023. mahanlp: A marathi natural language processing library. *arXiv preprint arXiv:2311.02579*.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2025. bytesizedllm@nlu of devanagari script languages 2025: Language identification using customized attention bilstm and xlm-roberta base embeddings. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi. International Committee on Computational Linguistics (ICCL).

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. Ethos: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, 8(6):4663–4678.

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

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. *arXiv preprint arXiv:1908.11049*.

Sanika Patil, Shraddha Nandvikar, Aakash Pardeshi, and Swapnali Kurhade. 2024. Automatic devanagari text summarization for youtube videos. In *2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)*, pages 16–21. IEEE.

Jebish Purbey, Siddartha Pullakhandam, Kanwal Mehreen, Muhammad Arham, Drishti Sharma, Ashay Srivastava, and Ram Mohan Rao Kadiyala. 2025. 1-800-shared-tasks@nlu of devanagari script languages 2025: Detection of language, hate speech, and targets using llms. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi. International Committee on Computational Linguistics (ICCL).

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*.

Durga Prasad Manukonda Rohith Gowtham Kodali and Daniel Iglesias. 2025. bytesizedllm@nlu of devanagari script languages 2025: Hate speech detection and target identification using customized attention bilstm and xlm-roberta base embeddings. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi. International Committee on Computational Linguistics (ICCL).

Shubham Shakya, Saral Sainju, Subham Krishna Shrestha, Prekshya Dawadi, and Shreya Khatiwada. 2025. Skpd emergency @ nlu of devanagari script languages 2025: Devanagari script classification using cbow embeddings with attention-enhanced bilstm. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi. International Committee on Computational Linguistics (ICCL).

Anik Mahmud Shanto, Mst. Sanjida Jamal Priya, and Mohammad Shamsul Arefin. 2025. Anisan@nlu of devanagari script languages 2025: Optimizing language identification with ensemble learning. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi. International Committee on Computational Linguistics (ICCL).

Arpit Sharma and BN Mithun. 2023. Deep learning character recognition of handwritten devanagari script: A complete survey. In *2023 IEEE International Conference on Contemporary Computing and Communications (InC4)*, volume 1, pages 1–6. IEEE.

Deepawali Sharma, Aakash Singh, and Vivek Kumar Singh. 2024. Thar-targeted hate speech against religion: A high-quality hindi-english code-mixed dataset with the application of deep learning models for automatic detection. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Rushendra Sidibomma, Pransh Patwa, Parth Patwa, Aman Chadha, Vinija Jain, and Amitava Das. 2025. Llmsagainsthate@nlu of devanagari script languages 2025: Hate speech detection and target identification in devanagari languages via parameter efficient fine-tuning of llms. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi. International Committee on Computational Linguistics (ICCL).

Akshay Singh and Rahul Thakur. 2024. Generalizable multilingual hate speech detection on low resource indian languages using fair selection in federated learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7204–7214.

Sukhjinder Singh, Naresh Kumar Garg, and Munish Kumar. 2023. Feature extraction and classification techniques for handwritten devanagari text recognition: a survey. *Multimedia Tools and Applications*, 82(1):747–775.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

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.

Anish Thapaliya, Prabhat Ale, and Suman Paudel. 2025. Mdsbots@nlu of devanagari script languages 2025: Detection of language, hate speech, and targets using murtweet. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi. International Committee on Computational Linguistics (ICCL).

Sulav Timilsina, Milan Gautam, and Binod Bhattarai. 2022. Nepberta: Nepali language model trained in a large corpus. In *Proceedings of the 2nd conference of the Asia-pacific chapter of the association for computational linguistics and the 12th international joint conference on natural language processing*. Association for Computational Linguistics (ACL).

Farjana Alam Tofa, Lorin Tasnim Zeba, Md Osama, and Ashim Dey. 2025. Cuet_insights@nlu of devanagari script languages 2025: Leveraging transformer-based models for target identification in hate speech. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi. International Committee on Computational Linguistics (ICCL).

Abhishek Velankar, Hrushikesh Patil, Amol Gore, Shubham Salunke, and Raviraj Joshi. 2022. L3cube-mahahate: A tweet-based marathi hate speech detection dataset and bert models. *arXiv preprint arXiv:2203.13778*.

Ashok Yadav and Vrijendra Singh. 2025. Dll5143a@nlu of devanagari script languages 2025: Detection of hate speech and targets using hierarchical attention network. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, Abu Dhabi. International Committee on Computational Linguistics (ICCL).

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

A Related Works

Identifying Devanagari script in social media is an increasingly challenging task that requires the attention of scholars, policymakers, and society (Sharma and Mithun, 2023; Singh et al., 2023). Hate speech detection has been a prominent research domain, with numerous studies concentrating on English and other widely spoken languages (Basile et al., 2019; Ousidhoum et al., 2019). However, efforts to identify hate speech in Devanagari-script languages, including Hindi, Nepali, and Marathi, are still insufficient (Sharma et al., 2024; Velankar et al., 2022; Singh and Thakur, 2024). Few studies have made substantial progress in identifying the targets of hate speech. Current studies generally expand hate speech classification to identify targets as persons or groups based on established criteria (Mathew et al., 2021; Mollas et al., 2022). Nevertheless, these approaches often rely primarily on English-centric models, with minimal adaptations for Devanagari script. Few studies have begun to fill the gap by leveraging multilingual embeddings to identify Devanagari script and its applications (Magdum et al., 2023; Timilsina et al., 2022; Gupta et al., 2022). Despite these advancements, the domain continues to encounter obstacles due to the varied linguistic attributes of the Devanagari script and the restricted availability of high-quality labeled datasets. To overcome the problem of reliable annotated datasets, researchers curated the corpus specifically focused on Devanagari languages like Hindi, Marathi, Bhojpuri, Sanskrit and Nepali (Jafri et al., 2024; Kulkarni et al., 2021; Ojha, 2019; Aralikatte et al., 2021; Rauniyar et al., 2023). This shared task utilizes the Devanagari dataset to engage scholars and professionals in addressing the problem of language identification, hate speech, and its target identification in the corpus.

B Evaluation and Competition

This section explains the nature of our competition, including the system for calculating rankings and other important details.

B.1 Evaluation Metrics

We employed accuracy, precision, recall, and macro F1-score to evaluate the performance. The macro F1-score sorting method was used to establish the participants' rank.

B.2 Competition Setup

We used the Codalab[1] to organize our competition. There were two stages to the competition: an evaluation stage where participants were introduced to the Codalab system, and a testing phase where the ultimate leaderboard ranking was established based on performance.

Registration: The shared task attracted 113 participants. Our shared task had interest from a diverse range of backgrounds and regions as anticipated by the email domains they registered with. Of these, 32 teams submitted their predictions.

Competition Timelines: On August 19, 2024, participants received access to the training and evaluation data, marking the beginning of the task. This initial phase aimed to help participants become familiar with the dataset and task requirements. The test phase started on September 27, 2024, when test data was made available without ground truth labels. Originally scheduled to end on October 17, 2024, the testing period was extended to October 27, 2024, in response to requests from participants, allowing additional time to complete submissions. The deadline for system description paper submissions was also extended from November 3 to November 10, 2024, providing more time for participants to document their methods. This structured timeline allowed participants to fully engage with each phase. We also ensured that support was provided to the participants in case of technical difficulties.

[1]The competition page can be found here: https://codalab.lisn.upsaclay.fr/competitions/20000.