

TeamHerald@CHIPSAL 2026: Hate Speech Detection and Sentiment Analysis of Nepali Memes using Transformer-based Architectures and Ensemble Learning

Ashish Acharya*, Anish Khatiwada*, Rohit Khadka, Pragya Aryal

Herald College Kathmandu, Kathmandu, Nepal

{ashish.acharya048, anishkhatioda, rohitkhadka153, pragyaaryal18}@gmail.com

*These authors contributed equally to this work.

Abstract

The analysis of internet memes in the Nepali language is complicated by frequent code-mixing and a lack of established baseline resources. While memes inherently combine visual and textual elements, this study focuses on a text-centric approach by extracting embedded text using an OCR layer and modeling it with Transformer-based architectures. We evaluate six distinct models and investigate the comparative effectiveness of Hard and Soft Voting ensemble strategies across two tasks: binary hate speech detection and three-class sentiment analysis. Experimental results show that a standalone decoder-only model achieved the highest performance for binary classification, whereas the Soft Voting ensemble performed best for the multi-class sentiment task, yielding a 15.8% relative improvement in Macro F1-score over the strongest standalone baseline. These findings suggest that ensemble strategies behave differently across binary and multi-class tasks, highlighting the importance of selecting aggregation methods suited to the classification objective.

Keywords: Nepali Language Processing, Meme Analysis, Ensemble Methods, Code-Mixed Text

1. Introduction

Internet memes have evolved from simple units of humor into sophisticated instruments for socio-political discourse, cultural expression, and, increasingly, the dissemination of harmful content. Unlike traditional text-only social media posts, memes are inherently multimodal, with meaning emerging from the interaction between visual imagery and embedded text (Kiela et al., 2020). This combination often creates compositional effects, where the sentiment of the text can be inverted or intensified by the accompanying image, rendering unimodal detection systems largely ineffective (Pramanick et al., 2022).

For low-resource languages such as Nepali, these challenges are compounded. The linguistic landscape features frequent code-mixing, including Devanagari and Romanized Nepali alongside English, which complicates standard tokenization and semantic analysis (Rauniyar et al., 2023). In addition, there is a scarcity of high-quality, annotated multimodal datasets that capture the cultural and political nuances of Nepali digital content. While significant progress has been made in English-centric meme analysis, such models often fail to generalize to South Asian contexts, where hate speech and sentiment are closely tied to local events, ethnic identities, and region-specific political satire.

To address these limitations, this paper evaluates a robust text-centric pipeline for Nepali meme analysis. An Optical Character Recognition (OCR) layer extracts textual content from raw images, simulat-

ing a real-world scenario in which ground-truth text is unavailable. We then benchmark six Transformer-based architectures, including models pre-trained on Nepali and cross-lingual corpora, to identify the most effective base learners for this domain.

The main contribution of this work is a systematic evaluation of ensemble strategies within a multimodal framework. By applying both Hard (majority-based) and Soft (probability-based) voting across binary and multi-class tasks, we examine how ensemble logic interacts with task complexity. Our analysis shows that Hard Voting stabilizes binary hate speech detection by mitigating model-specific noise, while Soft Voting is more effective for multi-class sentiment classification. This indicates that preserving the probabilistic confidence of underlying models is beneficial for resolving subtle linguistic ambiguities in multi-class social media discourse.

2. Related Work

Significant advances have been made in hate speech detection and sentiment analysis, particularly for high-resource languages such as English. However, low-resource languages, including Nepali, remain underrepresented in both dataset availability and model development (Parihar et al., 2021). The challenge is further amplified in the context of multimodal content such as memes, where meaning emerges from the interaction between textual and visual modalities.

2.1. Hate Speech and Sentiment Analysis in Low-Resource Languages

Hate speech and sentiment analysis in low-resource languages has gained increasing attention due to the rapid expansion of South Asian social media content. However, compared to English, languages such as Nepali, Hindi, Bangla, and Hinglish present significant challenges due to limited annotated datasets, linguistic diversity, and frequent code-mixing.

Early sentiment analysis work in Nepali relied on traditional machine learning models such as Support Vector Machines (SVM) and Naive Bayes with hand-crafted features (Sitaula et al., 2021). These approaches demonstrated limited generalization capability. The emergence of deep learning enabled the use of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, which improved performance through contextual embeddings. (Joshi et al., 2019)

For Bangla and Hindi, several studies have investigated abusive and hate speech detection in social media, emphasizing issues such as informal spelling, transliteration, and cultural nuance (Das and Mukherjee, 2021). Transformer-based models, particularly multilingual and monolingual pretrained architectures, have consistently demonstrated improved performance in these settings (Rathore and Chaurasia, 2025; Das et al., 2022).

However, recent findings suggest that domain-relevant non-hateful pretraining data may yield performance improvements comparable to explicitly hateful pretraining, challenging assumptions about task-specific pretraining strategies (Gokhale et al., 2022). Despite progress, data scarcity and domain adaptation remain core challenges in low-resource setups.

2.2. Multimodal Hate and Meme Analysis

With memes becoming a dominant medium for political expression and social commentary, hate speech detection has expanded beyond text-only analysis to multimodal settings. Memes often encode meaning through the interaction of visual and textual components, making unimodal models insufficient (Ji et al., 2023).

Previous multimodal hate detection research shows that models integrating both visual and textual features outperform text-only approaches (Shah et al., 2024). In many cases, hateful or sarcastic intent is not explicit in either modality alone but emerges from their combination (Yadav et al., 2024). Election-related memes, particularly in South Asian contexts, have been shown to propagate hate and misinformation, requiring culturally informed modeling approaches (Jafri et al., 2024).

Prompt-based and contrastive learning methods

have recently gained attention for modeling implicit cross-modal relationships without requiring extensive labeled datasets (Ranasinghe et al., 2021). CLIP-based architectures and adapter-based fine-tuning strategies further enhance multimodal meme classification, though detecting creative, humorous, or sarcastic hateful memes remains an open challenge, especially in multilingual and low-resource environments (Shah et al., 2024; Ji et al., 2023).

2.3. Datasets and Annotation Challenges

A major bottleneck in low-resource hate speech and sentiment analysis is the scarcity of high-quality annotated datasets. Creating datasets for languages such as Nepali, Hindi, and Bangla requires linguistic expertise and cultural understanding, particularly when addressing political or election-related discourse (Jafri et al., 2024).

Multimodal meme datasets introduce additional complexity, as annotators must interpret visual context, textual meaning, humor, sarcasm, and socio-political nuance simultaneously (Ji et al., 2023; Shah et al., 2024). Many existing South Asian hate speech datasets remain text-only and relatively small in scale (Das and Mukherjee, 2021; Gokhale et al., 2022).

Code-mixed datasets further complicate annotation due to ambiguous language boundaries, inconsistent spelling, and annotator disagreement (Rathore and Chaurasia, 2025). Studies frequently report moderate inter-annotator agreement for subjective labels such as hate and sarcasm (Ranasinghe et al., 2021). These challenges highlight the need for clearer annotation protocols and scalable dataset creation strategies in low-resource multimodal research.

2.4. Modeling Approaches

Modeling strategies for hate speech and sentiment classification have evolved from traditional bag-of-words and TF-IDF approaches to deep neural architectures (Sitaula et al., 2021; Das and Mukherjee, 2021). While traditional models are computationally efficient, they struggle to capture semantic and contextual nuance in informal social media text.

Recurrent neural networks (RNNs), LSTMs, and GRUs improved contextual modeling capabilities, particularly in Nepali and Hindi sentiment analysis (Joshi et al., 2019). Transformer-based architectures further advanced the field through large-scale pretraining and cross-lingual transfer learning (Rathore and Chaurasia, 2025; Das et al., 2022).

In multimodal scenarios, vision-language transformer models such as CLIP have shown strong performance by learning joint image-text representations (Shah et al., 2024). Prompt-based learning

and lightweight adaptation methods offer promising directions for low-resource meme classification tasks (Ranasinghe et al., 2021). Nevertheless, interpretability and cultural sensitivity of large transformer models in low-resource contexts remain underexplored.

Despite substantial progress, several gaps persist. First, limited work simultaneously addresses low-resource languages, code-mixing, and multimodal meme analysis within a unified framework. Second, Nepali multimodal meme datasets remain scarce, restricting comprehensive evaluation. Third, although transformer and CLIP-based models demonstrate strong empirical performance, their interpretability and socio-cultural alignment in South Asian contexts require further investigation. Finally, scalable annotation strategies and weakly supervised or prompt-based methods are necessary to reduce reliance on large labeled datasets.

Addressing these gaps is critical for developing robust hate speech and sentiment classification systems for multilingual, low-resource, and multimodal social media environments.

3. Datasets and Tasks Description

The shared task consists of two different subtasks: Subtask A aims to detect hate speech in Nepali meme images, Subtask B is related to classifying the sentiment from Nepali meme images. For all of the mentioned subtasks, datasets were provided by the organizers, which were initially created and curated by different papers (Thapa et al., 2026; Sarveswaran et al., 2026).

3.1. Sub-Task A

It involves binary classification to distinguish between images classified as HATE (labeled as 1) and NO-HATE (labeled as 0) (Thapa et al., 2025a,b; Bhandari et al., 2023). For the training purposes of this sub-task, a total of 1068 datasets were provided, out of which 720 were labeled as Hate and 348 were labeled as No Hate. The other 133 datasets were provided for validation, and 134 for testing purposes.

3.2. Sub-Task B

It involves multi-class classification to distinguish between images classified as Negative (labeled as 0), Neutral (labeled as 1), and Positive (labeled as 2) (Thapa et al., 2025a,b; Bhandari et al., 2023). For the training purpose of this sub-task, a total of 1061 datasets were provided, out of which 473 were labeled as Negative, 341 were labeled as Neutral and 247 were labeled as Positive. The other 133 data sets were for validation, and 133 for testing purposes.

4. Methodology

Our approach was designed to navigate the inherent multimodality of memes, where visual and textual cues are often inextricably linked. We developed a pipeline that first extracts embedded text via EasyOCR and subsequently processes these strings through a Transformer-based ensemble. While ground-truth text labels were available during the training phase, we intentionally restricted the testing phase to raw images. This design choice was made to simulate a real-world deployment scenario where OCR noise and extraction failures are inevitable (Rauniyar et al., 2023).

4.1. Dataset Balancing

A major challenge in our data pipeline was class imbalance, which can bias models toward majority labels and negatively impact Macro F1-score. To address this, we applied random oversampling to the minority classes to obtain a more balanced training distribution. Specifically, samples were resampled with replacement to reach $n = 1000$ instances per class for Subtask B and $n = 1500$ for Subtask A.

These target sizes were selected based on preliminary experiments, where increasing the representation of minority classes led to improved validation performance, particularly for Subtask A. This approach helps ensure that all classes contribute more evenly to the loss function during training, reducing bias toward majority classes.

4.2. Model Selection

To capture code-mixed and low-resource linguistic features, we fine-tuned six Transformer-based architectures from Hugging Face, each selected for specific advantages:

- **NepaliBERT**: BERT-base model specialized for native Nepali semantics, providing a robust baseline for formal and semi-formal Devanagari text.
- **DistilBERT-Base-Nepali**: Knowledge distilled, computationally efficient variant for Nepali text, included to evaluate performance trade-offs in resource-constrained environments.
- **XLM-RoBERTa-Base**: Multilingual model capable of handling complex code-mixed content by leveraging shared representations across 100 languages.
- **RoBERTa-Hindi**: Used for cross-lingual transfer learning; shared Devanagari script and lexical overlap between Hindi and Nepali may en-

hance feature representation for low-resource tasks.

- **distilgpt2-nepali**: Decoder-only GPT-2 model with continual pre-training on 13 million Nepali sequences. This model was evaluated only as a standalone baseline and was not included in the ensemble.

4.3. Experimental Setup

All models were fine-tuned using the parameters detailed in Table 1. We implemented **Early Stopping** with a patience of 3 epochs and a threshold of 0.01 to monitor validation loss.

Table 1: Fine-Tuning Hyperparameters

Hyperparameter	Value
Epochs	10
Batch Size	16
Learning Rate	2×10^{-5}
Weight Decay	0.01
Early Stopping Patience	3
Logging Steps	50

4.4. Evaluation Framework: Individual and Ensemble Testing

Following the evaluation of the six standalone models, we implemented an ensemble layer to aggregate their predictions. This was conducted using two distinct voting strategies applied across both subtasks to identify the most robust method for final classification.

4.4.1. Hard Voting

In the Hard Voting setup, the final prediction \hat{y} is derived from a majority consensus of the five base learners. In this configuration, each model contributes a single discrete label, and the most frequent class is selected as the final output. This method treats all models as equal contributors regardless of their individual confidence scores.

4.4.2. Soft Voting

We concurrently tested a Soft Voting strategy, which aggregates the predicted class probabilities (softmax outputs) rather than discrete labels. The final predicted class \hat{y} is calculated as:

$$\hat{y} = \arg \max_i \sum_{j=1}^5 w_j p_{ij}$$

where p_{ij} represents the probability assigned to class i by model j . This allows the ensemble to

account for the relative confidence of each base learner across the 2-class (Subtask A) and 3-class (Subtask B) problems.

5. Results and Analysis

The quantitative performance across both subtasks, summarized in Table 2, highlights architectural differences in low-resource multimodal NLP settings. In accordance with the official shared task evaluation criteria, Macro F1-score is reported as the primary metric.

In Subtask A (Hate Speech), the standalone Sakonii/distilgpt2-nepali model achieved the highest performance ($F1 = 0.6550$). This result may indicate that domain-specific continual pre-training contributed to improved binary discrimination between offensive and non-offensive content. Among the encoder-based models, RoBERTa-Hindi and NepaliBERT achieved identical evaluation metrics across all reported measures; however, their predictions differ at the instance level, suggesting similar class-wise error distributions. Although the ensemble approaches (Hard and Soft Voting) remained competitive, they did not surpass the GPT-2 baseline. This suggests that for binary classification tasks, a well-adapted single model can perform comparably to or better than aggregation-based methods.

In contrast, Subtask B (Sentiment Classification) revealed greater benefits from model aggregation under a three-class setting. The Soft Voting ensemble achieved the highest Macro F1-score (0.5518), representing a 15.8% relative improvement over the strongest standalone baseline (RoBERTa-Hindi). This pattern suggests that probabilistic aggregation may better capture uncertainty in multi-class sentiment classification compared to majority-based voting. Overall, these findings indicate that in low-resource settings, performance gains may depend not only on model selection but also on the aggregation strategy employed.

6. Limitations

While the proposed approach demonstrates performance gains for Nepali meme classification, several limitations should be considered when interpreting these findings.

First, the evaluation relies on a text-centric pipeline dependent on the EasyOCR extraction layer. Because the classification models operate solely on extracted text without incorporating visual features, overall performance is sensitive to OCR quality. A qualitative inspection of OCR outputs revealed frequent errors in stylized fonts, low-resolution text, and visually complex layouts, often resulting in incomplete or distorted tokens. These

Table 2: Experimental Results: Comparative Performance of Standalone Base Learners and Ensemble Strategies across Subtasks A and B.

Task	Model/Strategy	F1 Macro	Accuracy	Precision	Recall
Subtask A (Hate Speech)	Twitter-XLM-RoBERTa-Base	0.5830	0.6343	0.5836	0.5826
	DistilBERT-Base-Nepali	0.5318	0.5970	0.5332	0.5316
	RoBERTa-Hindi	0.5878	0.6343	0.5874	0.5884
	NepaliBERT	0.5878	0.6343	0.5874	0.5884
	XLM-RoBERTa-Base	0.5387	0.6343	0.5545	0.5419
	Sakonii/distilgpt2-nepali	0.6550	0.7090	0.6652	0.6497
	Hard Voting (Ensemble)	0.6422	0.6940	0.6482	0.6386
	Soft Voting (Ensemble)	0.6310	0.6866	0.6382	0.6273
Subtask B (Sentiment)	Twitter-XLM-RoBERTa-Base	0.4110	0.4286	0.4456	0.4333
	DistilBERT-Base-Nepali	0.4341	0.4586	0.4341	0.4352
	RoBERTa-Hindi	0.4764	0.5789	0.6481	0.4918
	NepaliBERT	0.3945	0.3985	0.3955	0.4083
	XLM-RoBERTa-Base	0.4484	0.4511	0.4484	0.4623
	Sakonii/distilgpt2-nepali	0.4108	0.4436	0.4222	0.4077
	Hard Voting (Ensemble)	0.4622	0.4887	0.5082	0.4635
	Soft Voting (Ensemble)	0.5518	0.5639	0.5659	0.5452

issues may lead to error propagation, where downstream classification is performed on noisy or partially incorrect inputs.

Second, the linguistic complexity of the dataset remains a persistent challenge. The prevalence of code-mixing (Nepali, English, and Romanized Nepali), along with colloquial expressions, may not be fully captured by existing pre-training corpora. This limitation is particularly evident in the three-class sentiment task, where subtle contextual and semantic distinctions are required to differentiate between neutral and polarized classes.

Finally, the use of random oversampling to balance class distributions creates a controlled training setup that does not reflect the naturally skewed distribution of hate speech in real-world social media. As a result, the reported performance should be interpreted within the context of balanced experimental conditions rather than as a direct estimate of real-world deployment performance.

7. Conclusion

We evaluated several Transformer architectures and ensemble strategies for hate speech and sentiment analysis in the low-resource Nepali language. The experimental results suggest differing strengths between standalone and ensemble architectures across tasks. While a standalone decoder-only model (Sakonii/distilgpt2-nepali) achieved the highest performance in binary hate speech detection, the Soft Voting ensemble achieved the strongest performance in the three-class sentiment task. Specifically, the 15.8% relative improvement in F1-Macro for Subtask B suggests that probabilistic aggregation may be beneficial for multi-class

sentiment classification.

These findings indicate that performance in this task varies with class cardinality and aggregation strategy. While memes are inherently multimodal, this work adopts a text-centric approach based on OCR-extracted content, which may limit performance in cases where meaning is conveyed primarily through visual context. Future work will explore the integration of visual-semantic features, including models such as CLIP or Vision Transformers (ViT), to better capture image-text interactions. Additionally, the application of parameter-efficient fine-tuning techniques remains a promising direction to investigate the balance between the computational requirements of ensemble-based inference and classification accuracy.

8. Bibliographical References

- Aashish Bhandari, Siddhant B Shah, Surendra-bikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. Crisishatemm: Multimodal analysis of directed and undirected hate speech in text-embedded images from russia-ukraine conflict. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1994–2003.
- Mithun Das and Animesh Mukherjee. 2021. Banglaabusememe: A dataset for bengali abusive meme classification. In *Conference on Empirical Methods in Natural Language Processing*.
- Mithun Das, Punyajoy Saha, Binny Mathew, and Animesh Mukherjee. 2022. [HateCheckHIn](#):

- Evaluating Hindi hate speech detection models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5378–5387, Marseille, France. European Language Resources Association.
- Omkar Gokhale, Aditya Kane, Shantanu Patankar, Tanmay Chavan, and Raviraj Joshi. 2022. [Spread love not hate: Undermining the importance of hateful pre-training for hate speech detection](#).
- Farhan Jafri, Kritesh Rauniyar, Surendrabikram Thapa, Mohammad 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*.
- Junhui Ji, Wei Ren, and Usman Naseem. 2023. [Identifying creative harmful memes via prompt based approach](#). In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 3868–3872, New York, NY, USA. Association for Computing Machinery.
- Aditya Joshi, Ashutosh Upadhyay, and Shishir Kumar. 2019. Deep learning for Nepali sentiment analysis. *arXiv preprint arXiv:1911.00000*.
- Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 2611–2624.
- 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.
- Shraman Pramanick, Shivam Sharma, Nikolaos Dimaridis, Akhtar Md Shad, Tanmoy Chakraborty, Preslav Nakov, and Dinesh Manocha. 2022. Momenta: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4439–4455.
- T. Ranasinghe, E. Zampieri, and M. Mihaela. 2021. Multilingual and multimodal approaches for hate speech detection. In *SemEval Workshop*.
- B. S. Rathore and S. Chaurasia. 2025. [Fine tuning large language models for hate speech detection in hinglish and code mixed custom dataset through a socially responsible approach for safer digital platforms](#). *Discover Sustainability*, 6:1409.
- Krishna Rauniyar, Gaurav Paudel, Biraj Khanal, and Bal Krishna Bal. 2023. Detection of hate speech, targets, humor, and stance in Nepali memes. In *Proceedings of the 7th Workshop on Online Abuse and Harms (WOAH)*, pages 154–165. Association for Computational Linguistics.
- Kengatharaiyer Sarveswaran, Surendrabikram Thapa, Ashwini Vaidya, Tafseer Ahmed Khan, and Bal Krishna Bal. 2026. Findings of the second workshop on challenges in processing south asian languages (chipsal 2026). In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHI PSAL)*.
- Siddhant Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. 2024. [Memecclip: Leveraging clip representations for multimodal meme classification](#).
- C. Sitaula, A. Basnet, A. Mainali, and {T. B.} Shahi. 2021. Deep learning-based methods for sentiment analysis on nepali text. *Computational Intelligence and Neuroscience*.
- Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Kritesh Rauniyar, Laxmi Thapa, Surabhi Adhikari, Kristina T. Johnson, Kengatharaiyer Sarveswaran, Bal Krishna Bal, and Usman Naseem. 2026. Multimodal hate and sentiment understanding in low-resource text-embedded images for online safety and digital well-being. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHI PSAL)*.
- Surendrabikram Thapa, Hariram Veeramani, Liang Hu, Qi Zhang, Wei Wang, and Usman Naseem. 2025a. A multimodal prompt-based framework for analyzing code-mixed and low-resource memes. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 1913–1923.
- Surendrabikram Thapa, Hariram Veeramani, Imran Razzak, Roy Ka-Wei Lee, and Usman Naseem. 2025b. Cross platform multimodal retrieval augmented distillation for code-switched content understanding. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2042–2051.
- Anjali Yadav, Tanya Garg, Matej Klemen, Matej Ulcar, Basant Agarwal, and Marko Robnik Sikonja. 2024. [Code-mixed Sentiment and Hate-speech Prediction](#).