Multimodal Hate and Sentiment Understanding in Low-Resource Text-Embedded Images for Online Safety and Digital Well-being

Surendrabikram Thapa[1], Shuvam Shiwakoti[1, †], Siddhant Bikram Shah[2, ✪], Kritesh Rauniyar[3, §], Laxmi Thapa[4], Surabhi Adhikari[5], Kristina T. Johnson[2], Kengatharaiyer Sarveswaran[6], Bal Krishna Bal[7], and Usman Naseem[3]

[1]Virginia Tech, USA, [2]Northeastern University, USA, [3]Macquarie University, Australia,
[4]O.P. Jindal Global University, India, [5]Columbia University, USA
[6]University of Jaffna, Sri Lanka, [7]Kathmandu University, Nepal
[†]shuvam@vt.edu; [✪]shah.siddhantb@northeastern.edu; [§]rauniyark11@gmail.com

Abstract

This paper presents an overview of the Shared Task on Multimodal Hate and Sentiment Understanding in Low-Resource Memes, organized as part of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) at LREC 2026. The task addresses automated content understanding in low-resource settings by focusing on monolingual Nepali memes written in Devanagari script. Built upon the NeMeme dataset, the task comprises two subtasks: (1) binary hate speech detection and (2) three-class sentiment analysis. The competition attracted 23 teams for hate detection and 13 teams for sentiment analysis. Participating teams employed diverse strategies, including late-fusion multimodal architectures combining multilingual text encoders with vision models, caption-based approaches using large vision-language models, and ensemble techniques. The top-performing system achieved macro-F1 scores of 80.52% on hate detection and 68.81% on sentiment analysis using a late-fusion hybrid architecture with discriminative learning rates. Our analysis reveals that multimodal fusion consistently outperforms unimodal baselines, sentiment analysis poses greater challenges than hate detection due to increased semantic nuance, and the scarcity of Devanagari-centric pretrained models remains a significant bottleneck. This shared task establishes a benchmark for multimodal understanding in low-resource South Asian languages and provides insights for developing inclusive content moderation systems.

Keywords: Hate Speech Detection, Sentiment Analysis, Low-Resource Languages, Nepali Language, Vision-Language Models, Devanagari Script

1. Introduction

The rapid proliferation of social media has fundamentally transformed how individuals express opinions, disseminate ideas, and engage in public discourse across linguistic and cultural boundaries (Bhandari et al., 2023; Shiwakoti et al., 2024). Among the diverse forms of digital communication, memes–multimodal artifacts that interweave images and embedded text–have emerged as one of the most pervasive and influential vehicles for conveying humor, political commentary, social criticism, and cultural identity (Shifman, 2013). Their virality and accessibility make them powerful instruments for shaping public sentiment, but they also serve as conduits for hate speech, misinformation, and targeted abuse against individuals and communities (Pramanick et al., 2021). The inherent ambiguity of memes, where the boundary between satire and genuine malice is often deliberately blurred, presents a formidable challenge for both human moderators and automated content moderation systems (Shah et al., 2024; Shang et al., 2021).

While significant progress has been made in the computational analysis of memes in high-resource languages such as English (Kiela et al., 2020), the overwhelming majority of the world's languages remain critically underserved by existing tools and resources. This disparity is particularly consequential for languages that are widely spoken yet lack the annotated datasets, pretrained models, and benchmark evaluations necessary to drive progress in multimodal understanding. Nepali, a Devanagari-script language spoken by over 20 million people worldwide and the official language of Nepal, exemplifies this challenge. Despite the strong presence of politically and socially driven memes in Nepali online spaces, computational resources for understanding such content remain scarce (Thapa et al., 2025a). The low-resource nature of Nepali, compounded by its rich morphological structure, cultural specificity, and the prevalence of code-mixed expressions in digital communication, renders the direct application of models developed for high-resource languages both unreliable and culturally inadequate.

The challenge of analyzing Nepali memes is amplified by several interrelated factors. First, the multimodal nature of memes requires systems to jointly reason over visual and textual modalities–a task complicated by the fact that dominant vision-language models such as CLIP (Radford et al., 2021) are predominantly pretrained on English-centric web data and exhibit limited representa-

Proceedings of the Second workshop on Challenges in Processing South Asian Languages (CHiPSAL2026) @ LREC, pages 219–228
16 May 2026.

tional capacity for Devanagari script and culturally specific South Asian visual content (Poudel et al., 2025). Second, the nuanced interplay between humor, sarcasm, political satire, and genuine hate in Nepali memes demands models that can disentangle layered communicative intents rather than relying on surface-level lexical or visual cues (Rai et al., 2023). Third, the scarcity of annotated data imposes severe constraints on model training, necessitating innovative strategies for data augmentation, cross-lingual transfer, and efficient fusion of multimodal signals (Thapa et al., 2025a).

To address these critical research gaps, we present the **Shared Task on Multimodal Hate and Sentiment Understanding in Low-Resource Memes**, organized as part of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) (Sarveswaran et al., 2026), co-located with LREC 2026. Building on the NeMeme dataset (Thapa et al., 2025a)–the first publicly available annotated dataset for hate speech and sentiment analysis in Nepali memes–this shared task is designed to catalyze the development of robust multimodal systems capable of operating in low-resource settings. The task encompasses two complementary subtasks: (1) **Hate Speech Detection**, a binary classification task requiring systems to distinguish hateful memes from non-hateful ones, and (2) **Sentiment Analysis**, a three-class classification task requiring systems to assign negative, neutral, or positive sentiment labels to memes. Both subtasks operate exclusively on monolingual Nepali memes written in the Devanagari script, evaluated using macro-averaged F1-score to account for inherent class imbalance.

The shared task attracted participation from a diverse set of teams employing a wide spectrum of methodological strategies, ranging from unimodal text-only and vision-only baselines to multimodal fusion architectures incorporating cross-modal attention, learnable gating mechanisms, and vision-language model-generated captions. Participants explored innovative approaches to circumvent the limitations of existing pretrained models for Devanagari content, including the use of large vision-language models such as Gemini for stochastic caption generation, multilingual encoders such as BGE-M3 and XLM-RoBERTa for text representation, and various fusion strategies, including early concatenation, late voting, weighted fusion, and hybrid attention-based architectures. The diversity of submitted approaches and their varying degrees of success provide valuable insights into the relative merits of different multimodal strategies under extreme data scarcity and linguistic resource constraints.

This paper provides a comprehensive overview of the shared task, including a description of the dataset and its annotation, the evaluation protocol and metrics for each subtask, a summary of the participating teams and their methodologies, and an analysis of the results. Through this shared task, we aim to advance the state of multimodal content understanding for low-resource South Asian languages and contribute to the development of more inclusive, culturally aware, and effective systems for online safety and digital well-being.

2. Related Works

Detecting hate speech on social media has become an increasingly important area of research, with most efforts concentrating on text-based content (Chhabra and Vishwakarma, 2023; Parihar et al., 2021). However, there has been a notable surge in scholarly interest in identifying hate speech in memes or text-containing images (Ji et al., 2023; Thapa et al., 2025b), as multimodal hate and sentiment analysis on social media remains an increasingly complex challenge for researchers, policymakers, and society (Jahan and Oussalah, 2023). Text-embedded images, including memes and other visual-text content like TV headline screenshots, combine images for context and text for information. While meme analysis has received significant attention, detecting hate speech in such text-embedded images remains underexplored, motivating this shared task.

Similarly, research on memes and multimodal textual-visual data has predominantly focused on general social media platforms, with an emphasis on the English language (Hossain et al., 2022; Xu et al., 2022; Shah et al., 2024). Efforts to create dedicated datasets for low-resource languages are quite limited, especially for the Nepali language, which is spoken by approximately 40 million people worldwide (Thapa et al., 2023; Poudel et al., 2025). A few studies have explored the Hindi language, where researchers curated multimodal datasets of memes and sarcastic content (Dubey et al., 2025; Kumari et al., 2024). However, in the context of the Nepali language and multimodal memes, there is a severe lack of datasets that comprehensively capture multiple aspects of social dynamics, such as cultural nuances, humor, sentiment, and hate speech, limiting research in this underexplored domain. Thapa et al. (2025a) introduced a unique dataset featuring multimodal Nepali and code-mixed Nepali memes (combining Nepali and English). These efforts are progressively laying the foundation for future research on language-specific multimodal content. This shared task aims to engage the research community and encourage their participation in context-specific investigations, specifically focusing on Nepali meme identification.

3. Shared Task Description

Our shared task targets the automated understanding of monolingual Nepali memes across two complementary subtasks: hate speech detection and sentiment analysis. Both subtasks operate exclusively on memes written in the Nepali language using the Devanagari script, posing unique challenges due to the low-resource nature of the language, its rich cultural context, and the inherently multimodal nature of meme content.

3.1. Subtask A: Hate Speech Detection

Hate Speech Detection in Nepali Memes: The aim of this subtask is to detect the presence of hate speech in monolingual Nepali memes. Given a meme, participating systems must classify it into one of two categories: *Non-Hate* and *Hate*. Hateful memes are defined as those that vilify, denigrate, bully, insult, or mock a subject based on characteristics such as gender, race, religion, caste, or organizational status, as well as memes that contain hateful symbols or glorify violence. Non-hateful memes, by contrast, may convey humor, affection, motivation, or constructive criticism without malicious intent. This subtask is evaluated using macro-averaged F1-score to account for the natural class imbalance between hate and non-hate instances in the dataset.

3.2. Subtask B: Sentiment Analysis

Sentiment Analysis in Nepali Memes: The goal of this subtask is to classify the sentiment expressed in monolingual Nepali memes. Given a meme, participating systems must assign one of three sentiment labels: *Negative*, *Neutral*, or *Positive*. Negative memes are those that aim to denigrate, insult, or belittle a subject based on their social, personal, or organizational status. Neutral memes present an objective perspective without conveying strong emotional tones. Positive memes express affection, support, gratitude, praise, or motivation. This subtask is evaluated using macro-averaged F1-score across all three sentiment classes.

4. Dataset

This section describes the dataset used in our shared task, including its source, annotation procedure, and the statistics of the train, validation, and test splits used for evaluation.

4.1. Source and Collection

We utilize a subset of the NeMeme dataset (Thapa et al., 2025a), the first publicly available annotated dataset for hate speech and sentiment analysis in Nepali memes. The full NeMeme dataset was collected from multiple social media platforms, including Twitter, Instagram, Reddit, Facebook, and Threads. For our shared task, we focus exclusively on the monolingual Nepali subset, which consists of memes written entirely in the Nepali language using the Devanagari script. Code-mixed memes combining Nepali and English were excluded from this task.

4.2. Annotation

Each meme in the dataset is annotated for two tasks: (1) binary hate speech detection (hate vs. non-hate) and (2) three-class sentiment analysis (negative, neutral, positive). Annotations were produced by three annotators proficient in Nepali, Hindi, and English, following a three-phase annotation scheme (Bhandari et al., 2023) comprising a pilot run, a revision phase, and a consolidation phase. Inter-annotator agreement, measured using Fleiss' κ, was 0.74 for hate speech and 0.69 for sentiment, indicating substantial agreement.

4.3. Data Splits

We partition the dataset into training, validation, and test sets for each task. Table 1 summarizes the label distributions across splits.

For the **hate speech detection** task, the training set contains 1,068 memes (348 hate, 720 non-hate), the validation set contains 133 memes (35 hate, 98 non-hate), and the test set contains 134 memes (44 hate, 90 non-hate). The dataset exhibits a natural class imbalance, with non-hateful memes comprising approximately 67% of instances across all splits, reflecting realistic distributions on social media platforms.

For the **sentiment analysis** task, the training set contains 1,061 memes (341 negative, 473 neutral, 247 positive), the validation set contains 133 memes (39 negative, 65 neutral, 29 positive), and the test set contains 133 memes (40 negative, 63 neutral, 30 positive). Neutral memes represent the plurality class, consistent with the broader NeMeme dataset statistics.

Table 1: Dataset statistics for the shared task across train, validation, and test splits.

Task	Label	Train	Val	Test	Total
Hate Speech	Hate	348	35	44	427
	Non-Hate	720	98	90	908
	Total	1068	133	134	1335
Sentiment	Negative	341	39	40	420
	Neutral	473	65	63	601
	Positive	247	29	30	306
	Total	1061	133	133	1327

Rank	Participant	Accuracy (%)	Recall (%)	Precision (%)	F1-score (%)
1	linus (Regmi et al., 2026)	82.09	82.02	79.74	80.52
2	ZeroR (Khanal, 2026)	82.09	79.70	79.70	79.70
3	eGrantha.ai (Thapaliya, 2026)	78.36	72.85	76.12	73.97
4	samirwagle (Wagle et al., 2026)	70.15	70.23	68.17	68.34
5	team Oryu (Orny et al., 2026)	73.13	65.48	69.65	66.40
6	EthosAI (Bansal et al., 2026)	71.64	65.53	67.43	66.14
7	ashweta (Fondekar et al., 2026)	71.64	64.95	67.41	65.64
8	anish (Acharya et al., 2026)	70.90	64.97	66.52	65.50
9	CUET-DoubleA	65.67	66.89	64.93	64.33
10	sundess	67.16	62.78	62.78	62.78
11	sameekxa	64.93	63.43	62.24	62.36
12	hasnat (Chowdhury et al., 2026)	68.66	61.57	63.47	62.02
13	researcher	67.91	61.01	62.58	61.41
14	sujata-gaihre	64.93	61.69	61.12	61.30
15	Multi-Modal-Minds (Shrestha et al., 2026)	61.19	59.49	58.61	58.52
16	deleted_user_52441	64.93	58.21	59.06	58.42
17	eshan	57.46	55.56	55.01	54.72
18	dutta_arka	58.21	49.14	48.96	48.57
19	The Argonauts	67.91	52.88	62.82	47.94
20	now_coder	64.18	48.36	41.54	40.99
21	akshayyy22	67.16	50.00	33.58	40.18
22	sujatagaihre	67.16	50.00	33.58	40.18
23	anycookie	38.81	46.31	45.86	38.59

Table 2: Sub-task A (Hate Speech Detection) Leaderboard, Ranked by Macro F1-score. All scores are presented as percentages (%). Note 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.

Rank	Participant	Accuracy (%)	Recall (%)	Precision (%)	F1-score (%)
1	linus (Regmi et al., 2026)	68.42	70.13	69.79	68.81
2	anish (Acharya et al., 2026)	56.39	54.52	56.59	55.18
3	sundess	71.43	60.91	50.67	54.81
4	ZeroR (Khanal, 2026)	52.63	53.57	52.68	51.77
5	samirwagle (Wagle et al., 2026)	60.15	53.48	55.23	51.12
6	sidratul (Muntaha et al., 2026)	51.13	52.01	50.70	50.45
7	dipika_nath (Debnath et al., 2026)	49.62	51.51	49.76	49.75
8	ashweta (Fondekar et al., 2026)	48.87	50.37	48.58	48.59
9	EthosAI (Bansal et al., 2026)	49.62	48.57	48.24	48.39
10	yinloonkhorr	49.62	47.10	47.89	47.26
11	researcher	44.36	40.77	42.22	41.08
12	Multi-Modal-Minds (Shrestha et al., 2026)	39.10	37.14	34.93	35.28
13	akshayyy22	45.11	36.59	36.85	33.36

Table 3: Sub-task B (Sentiment Analysis) Leaderboard, Ranked by Macro F1-score. All scores are presented as percentages (%). Note 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.

5. Evaluation and Competition

This section describes the structure of our competition, along with the methodology used to determine ranks and other relevant details.

5.1. Evaluation Metrics

To evaluate the effectiveness of the participants' contributions, we used four metrics: macro F1-score, accuracy, precision, and recall. The participants' final ranks were determined using the macro F1-score as the primary ranking metric.

5.2. Competition Setup

We used Codabench[1] [,2] to organize our competition. The competition was hosted as two separate competitions corresponding to the two subtasks. Each competition consisted of two phases: a development phase, where participants could familiarize themselves with the Codabench platform and de-

[1]Subtask A (Hate Detection): https://www.codabench.org/competitions/12090/

[2]Subtask B (Sentiment Analysis): https://www.codabench.org/competitions/12091/

velop their methods, and a test phase, where performance was used to determine the final ranking on the leaderboard. The results from the development phase were made available to participants after the phase concluded, enabling them to further refine their approaches for the test phase.

5.2.1. Registration

For Subtask A (Hate Detection), a total of 65 participants registered, out of which 23 teams submitted their predictions. For Subtask B (Sentiment Analysis), 43 participants registered, with 13 teams submitting their predictions.

5.2.2. Competition Timelines

The competition commenced on December 8, 2025, when training and development data were made available, marking the start of the development phase. During this phase, participants familiarized themselves with the Codabench platform and began developing their systems. The test phase began on December 25, 2025, when test data was provided without any ground truth labels. The test phase was originally scheduled to end on February 14, 2026; however, in response to requests from several participants, it was extended until February 22, 2026. The paper submission deadline was correspondingly updated from February 20, 2026 to March 1, 2026. Notification of acceptance was scheduled for March 20, 2026, with camera-ready papers due by March 30, 2026.

6. Participants' Methods

6.1. Overview

The leaderboards for Subtask A and Subtask B are presented in Table 2 and Table 3, respectively.

6.2. Methods

In the following subsections, we outline the approaches adopted by participating teams across both Subtasks A and B, as well as those specific to Subtask A and Subtask B individually, based on their system description papers.

6.2.1. Both Subtasks A and B

Linus (Regmi et al., 2026) proposed a late-fusion hybrid architecture combining CLIP ViT-B/32 for visual feature extraction with NepBERTa for textual encoding, targeting both Subtasks A and B. The extracted modality-specific embeddings are independently L2-normalised, concatenated into a 1280-dimensional joint representation, and passed through a multi-layer perceptron (MLP). Training employed discriminative learning rates, with substantially lower rates to prevent catastrophic forgetting. The training also used label smoothing, explicit class weighting for the underrepresented neutral class, and OneCycleLR scheduling. Their framework achieved Macro-F1 scores of 0.8052 and 0.6881 on Subtasks A and B, respectively, securing first position on both subtasks.

ZeroR (Khanal, 2026) introduced a two-stage vision-language system built on Qwen3-VL-8B-Instruct, leveraging the model's native Devanagari reading capability to process. In the first stage, LoRA adapters are applied across all linear projections for generative fine-tuning, accompanied by an MLP projection head mapping hidden states to a 512-dimensional embedding space. The second stage introduces supervised contrastive learning via InfoNCE loss jointly optimized with a linear classifier. The sentiment task demanded considerably stronger regularization, including lower LoRA rank, higher weight decay, and early stopping. The system achieved Macro-F1 scores of 0.7970 and 0.5177, securing second and fourth positions on Subtasks A and B, respectively.

MEME-Fusion (Wagle et al., 2026) proposed a hybrid cross-modal attention fusion architecture addressing both Subtasks A and B. The system combines CLIP ViT-B/32 for visual encoding with BGE-M3 for multilingual text representation, connected through 4-head self-attention and a learnable softmax-normalized gating network that dynamically weights modality contributions on a per-sample basis. Evaluation across eight model configurations demonstrated that the hybrid fusion model achieved a Macro-F1 of 0.6834 on Subtask A and 0.6102 on Subtask B, securing 4th and 5th place, respectively. The study also uncovered that English-centric vision models exhibited near-random performance on Devanagari script, and standard ensemble methods catastrophically degraded under data scarcity due to correlated overfitting.

EthosAI (Bansal et al., 2026) presented a multimodal approach for both Subtask A (Hate Speech Detection) and Subtask B (Sentiment Classification), systematically evaluating 10 image models and 10 text models before selecting the top two from each modality for fusion experiments. Text features were extracted using both EasyOCR and Qwen2-VL-7B-Instruct for image description generation, while four fusion strategies–simple concatenation, weighted fusion

with learnable scalars, cross-attention, and element-wise addition–were compared. For Subtask A, the best-performing configuration combined Sentence-Transformers/LaBSE for text and ResNet-18 for image using weighted fusion, achieving a Macro-F1 of 0.6614. For Subtask B, Sentence-Transformers/LaBSE with DeiT-Base using simple fusion achieved a Macro-F1 of 0.4839. Their experiments indicated that multimodal models consistently outperform unimodal baselines for meme understanding in low-resource settings.

Unigoa (Fondekar et al., 2026) investigated early-fusion versus late-fusion multimodal architectures for both Subtasks A and B. Their primary system employs a late-fusion dual-encoder design pairing XLM-RoBERTa with CLIP ViT-B/32, where text and image embeddings are concatenated and fed into an MLP classifier. For Subtask A, XLM-RoBERTa-Base is used alongside validation-based threshold tuning to handle class imbalance, while Subtask B adopts XLM-RoBERTa-Large with a suite of training stabilisation techniques including WeightedRandomSampler, Soft Focal Loss, gradient accumulation, and discriminative learning rates. As a comparative baseline, the authors also evaluated an early-fusion ViLT model with NepBERTa tokenization. The late-fusion approach consistently outperformed the early-fusion baseline, achieving Macro-F1 scores of 0.6564 and 0.4859 on Subtasks A and B, placing 7th and 8th respectively. Their analysis highlights the advantage of leveraging strong multilingual pretrained encoders in a late-fusion setup over joint vision–language transformers under low-resource data conditions, particularly for code-mixed memes containing both Devanagari and Roman-script text.

TeamHerald (Acharya et al., 2026) tackled both Subtasks A and B using a text-centric pipeline built on OCR-extracted text from meme images via EasyOCR, followed by fine-tuning six Transformer-based architectures including NepaliBERT, DistilBERT-Base-Nepali, XLM-RoBERTa-Base, Twitter-XLM-RoBERTa-Base, RoBERTa-Hindi, and a decoder-only distilgpt2-nepali model. Class imbalance was addressed through stratified oversampling prior to training. The authors systematically compared Hard Voting and Soft Voting ensemble strategies across both subtasks, revealing a task-dependent pattern: the standalone distilgpt2-nepali model achieved the best Macro-F1 of 0.6550 on Subtask A, while the Soft Voting ensemble yielded the highest Macro-F1 of 0.5518 on Subtask B, representing a 7.5% relative improvement over the best individual baseline. Their analysis suggests that probabilistic aggregation is more effective than majority voting for multi-class classification in low-resource settings, whereas binary tasks may benefit more from well-adapted standalone models.

Multi-Modal-Minds (Shrestha et al., 2026) conducted a broad comparative study spanning unimodal text models (mBERT, XLM-RoBERTa, MuRIL), unimodal visual models (ResNet, ConvNeXt, ViT), nine late-fusion combinations pairing each text encoder with each visual encoder, and the vision-language foundation model SigLIP. A stochastic image augmentation pipeline was used in all visual experiments to simulate web degradation, apply coarse dropout for spatial regularization, and introduce color variation. Surprisingly, instead of multimodal fusion performing best, the standalone ViT achieved the highest Macro-F1 on Subtask A (0.6278), while SigLIP performed best on Subtask B (0.5481). The architecture ranked 15th on Subtask A and 12th on Subtask B.

6.2.2. Subtask A

eGrantha.ai (Thapaliya, 2026) proposed a caption-based approach for hate speech detection using the Gemini family of models (Gemini 2.X and Gemini 3.X) to generate contextually rich captions from meme images, converting the multimodal problem into a standard text classification task. The system fine-tuned RoBERTa-base on the generated captions, employing a stochastic caption regeneration strategy to address class imbalance by doubling minority-class coverage. At inference time, Test-Time Augmentation (TTA) was applied using five independent captions generated by diverse Gemini model variants, with class-probability distributions averaged before argmax. This approach achieved a Macro-F1 of 0.7397, securing 3rd place on the official leaderboard, demonstrating that caption-based modeling with stochastic augmentation can compete with conventional multimodal fusion strategies.

team Oryu (Orny et al., 2026) only addressed Subtask A through a late-fusion multimodal framework pairing XLM-RoBERTa-Base for multilingual text encoding with a Vision Transformer (ViT-Base-Patch16-224) for image feature extraction. The 768-dimensional embeddings from each encoder are concatenated to form a 1536-dimensional joint representation. This representation is then processed by a two-layer fully connected classification head with ReLU activation and dropout regularization. The authors report an iterative development

process, progressing from a basic multimodal baseline (F1: 0.40) through a gated fusion variant (F1: 0.52) to the final late-fusion configuration. The system obtained a Macro-F1 of 66.40% on the official test set, ranking fifth on Subtask A.

HasNat (Chowdhury et al., 2026) tackled Subtask A by benchmarking three multimodal model families—ViT-B-32 (OpenCLIP), AltCLIP, and BLIP2+mT5—each evaluated under default, preprocessing, augmentation, and combined settings. All architectures employed a late-fusion strategy in which frozen vision–language backbones produced text and image embeddings that were concatenated and passed to a trainable MLP classifier. Training relied on 5-fold stratified cross-validation with weighted binary cross-entropy loss to handle class imbalance, and fold-level probability averaging for inference. Among the configurations explored, AltCLIP in its default setting yielded the highest validation Macro-F1 of 0.66, while the authors noted that preprocessing and augmentation effects were highly model-dependent, with aggressive augmentation sometimes degrading recall by distorting meme-embedded text.

6.2.3. Subtask B

NeuralNoodles (Muntaha et al., 2026) tackled Subtask B through a late-fusion multimodal stacking framework that integrates three complementary unimodal streams: a TF-IDF model combining word-level (1–2 gram) and character-level (3–5 gram) n-grams for lexical representation, a frozen `paraphrase-multilingual-MiniLM-L12-v2` Sentence Transformer for semantic encoding, and a fine-tuned EfficientNet-B0 for visual feature extraction. Each component produces a three-class probability distribution, which are concatenated into a 9-dimensional meta-feature vector and passed to a Logistic Regression meta-classifier trained via stratified 5-fold cross-validation. An ablation study revealed that while the image model was the strongest individual contributor (Macro-F1 of 0.3980), the full three-modality stack yielded a 6.4% relative improvement, achieving 0.4234 on cross-validation and 0.5045 on the official test set for 6th place. The authors note that two-modality combinations did not consistently outperform the best unimodal baseline, suggesting that complementary information across all three representation types is necessary for effective fusion in this low-resource setting.

Cuet Yet Another Baseline (Debnath et al., 2026) implemented a multimodal pipeline for sentiment Analysis that fuses three streams: contextual text features from a fine-tuned XLM-RoBERTa-large applied to OCR-extracted meme text combined with BLIP-2 captions, pooled visual features from a frozen CLIP ViT-L/14 encoder, and a BLIP-2 caption embedding along with a CLIP image–text cosine similarity scalar. The attended representation was concatenated with projected CLIP streams and passed through a two-layer MLP classifier with GELU activations and dropout. Code-mixed Nepali–English samples were incorporated as training augmentation, with monolingual samples upweighted via a WeightedRandomSampler. Their approach achieved a Macro-F1 of 0.50, securing 7th place.

7. Discussion

The results of this shared task reveal several important insights into multimodal hate speech and sentiment analysis for low-resource languages. The top-performing systems consistently demonstrated that late-fusion architectures pairing strong multilingual text encoders (NepBERTa, XLM-RoBERTa, BGE-M3) with established vision models (CLIP ViT) achieve superior performance compared to early-fusion or vision-language models like ViLT, suggesting that the scarcity of Devanagari-centric multimodal pretraining data makes separate encoding followed by learned fusion more effective than joint processing. Notably, the winning team's hybrid fusion with discriminative learning rates and careful regularization achieved macro-F1 scores of 80.52% and 68.81% on hate detection and sentiment analysis respectively, while caption-based approaches using large vision-language models like Gemini demonstrated competitive performance by converting the multimodal problem into text classification. The substantial performance gap between Subtask A (hate detection) and Subtask B (sentiment analysis), with the best F1 scores differing by approximately 12 percentage points, indicates that sentiment classification remains considerably more challenging, likely due to the inherent ambiguity in distinguishing neutral from mildly positive or negative memes and the greater semantic nuance required for three-way classification.

The diversity of methodological approaches adopted by participants underscores both the opportunities and challenges in low-resource multimodal understanding. While multimodal fusion generally outperformed unimodal baselines, the margin of improvement varied considerably across teams, with some fusion strategies yielding minimal gains or even degrading performance due to correlated overfitting under severe data scarcity. Vision-only models achieved surprisingly compet-

itive results on both subtasks, suggesting that meme images contain substantial visual cues independent of text, though text-centric models generally underperformed, highlighting the critical importance of visual context. The observed sensitivity to architectural choices, hyperparameter configurations, and augmentation strategies emphasizes the need for careful model design and extensive validation when working with low-resource languages. Future work should explore more sophisticated cross-lingual transfer learning, leverage larger multilingual vision-language models with improved Devanagari support, and investigate semi-supervised or self-supervised approaches to alleviate data scarcity constraints in Nepali and other underserved South Asian languages.

8. Conclusion

This paper presented an overview of the Shared Task on Multimodal Hate and Sentiment Understanding in Low-Resource Memes, organized as part of CHiPSAL 2026 at LREC 2026. The results demonstrate that late-fusion architectures combining multilingual text encoders with established vision models achieve the most robust performance, with the top system attaining macro-F1 scores of 80.52% on hate detection and 68.81% on sentiment analysis. Key insights include the consistent superiority of multimodal fusion over unimodal approaches, the greater difficulty of sentiment classification compared to hate detection, and the competitive performance of caption-based methods using large vision-language models. We hope this shared task catalyzes further research into multimodal understanding for low-resource South Asian languages and contributes to the development of more inclusive, culturally aware content moderation systems for underserved linguistic communities.

9. References

Ashish Acharya, Anish Khatiwada, Rohit Khadka, and Pragya Aryal. 2026. Teamherald@chipsal 2026: Hate speech detection and sentiment analysis of nepali memes using transformer-based architectures and ensemble learning. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*.

Vinayak Bansal, Deepawali Sharma, Aakash Singh, and Vivek Kumar Singh. 2026. Ethosai@chipsal2026: Hate and sentiment understanding in low-resource memes using a multimodal approach. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*.

Aashish Bhandari, Siddhant B Shah, Surendrabikram 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.

Anusha Chhabra and Dinesh Kumar Vishwakarma. 2023. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, 29(3):1203–1230.

Alvee Hasan Chowdhury, MD. Abul Hasnat, and Adnan Faisal. 2026. Hasnat@chipsal 2026: Multimodal hate speech detection in low-resource nepali memes using aligned vision–language models. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*.

Rotna Dipika Debnath, Shahrin Afroz Hoque Ruhi, Ayesha Labiba, Arpita Mallik, and Hasan Murad. 2026. Cuet yet another baseline@chipsal lrec 2026: Shared task on multimodal sentiment understanding in low-resource memes. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*.

Kriti Dubey, Vaishnavi Srivastava, Garima Sharma, Nonita Sharma, Deepak Sharma, Uttam Ghosh, Osama Alfarraj, and Amr Tolba. 2025. Multimodal detection of offensive content in hindi memes. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Ashweta Fondekar, Milind Shivolkar, and Jyoti Pawar. 2026. Unigoa@chipsal 2026: Early vs late fusion for multimodal hate and sentiment detection in nepali memes. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. Mute: A multimodal dataset for detecting hateful memes. 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: student research workshop*, pages 32–39.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, page 126232.

Junhui Ji, Wei Ren, and Usman Naseem. 2023. Identifying creative harmful memes via prompt based approach. In *Proceedings of the ACM web conference 2023*, pages 3868–3872.

Nitiz Khanal. 2026. Zeror@chipsal 2026: Two-stage vision-language adaptation with contrastive learning for nepali meme classification. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*.

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. *Advances in neural information processing systems*, 33:2611–2624.

Gitanjali Kumari, Chandranath Adak, and Asif Ekbal. 2024. Mu2sts: a mu ltitask mu ltimodal s arcasm-humor-differential t eacher-s tudent model for sarcastic meme detection. In *European Conference on Information Retrieval*, pages 19–37. Springer.

Sidratul Muntaha, Sabila Anzum, Arpita Mallik, and Hasan Murad. 2026. Neuralnoodles@chipsal 2026: Late-fusion multimodal stacking for nepali meme sentiment classification. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*.

Noore Tamanna Orny, Joyeta Barua Moni, and Md. Abtahee Kabir. 2026. Team oryu@chipsal 2026: Integrating text and vision transformers for multimodal hate speech detection in memes. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*.

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.

Sweta Poudel, Kritesh Rauniyar, Ashish Acharya, Junaid Rashid, Surabhi Adhikari, Usman Naseem, and Surendrabikram Thapa. 2025. Nepaes: exploring the promise of automated essay scoring for nepali essays. *PeerJ Computer Science*, 11:e3253.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Momenta: A multimodal framework for detecting harmful memes and their targets. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 4439–4455.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Pooja Rai, Sanjay Chatterji, and Byung-Gyu Kim. 2023. Deep learning-based sequence labeling tools for nepali. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(8):1–23.

Sunil Regmi, Bipesh Subedi, Saugat Singh, and Suman Shrestha. 2026. linus@chipsal 2026: Multimodal hate speech and sentiment detection in low-resource memes using late-fusion hybrid architecture. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*.

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 (CHiPSAL)*.

Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. 2024. Memeclip: Leveraging clip representations for multimodal meme classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17320–17332.

Lanyu Shang, Yang Zhang, Yuheng Zha, Yingxi Chen, Christina Youn, and Dong Wang. 2021. Aomd: An analogy-aware approach to offensive meme detection on social media. *Information Processing & Management*, 58(5):102664.

Limor Shifman. 2013. Memes in a digital world: Reconciling with a conceptual troublemaker. *Journal of computer-mediated communication*, 18(3):362–377.

Shuvam Shiwakoti, Surendrabikram Thapa, Kritesh Rauniyar, Akshyat Shah, Aashish Bhandari, and Usman Naseem. 2024. Analyzing the dynamics of climate change discourse on twitter: A new annotated corpus and multi-aspect classification. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, pages 984–994.

Sandesh Shrestha, Bikram K.C., Akshyat Shah, Ashish Acharya, and Rabin Thapa. 2026. Multimodal-minds@chipsal 2026: A comparative

study of textual, visual and multimodal architecture for nepali meme moderation. In *Proceedings of the Second 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. *Frontiers in Artificial Intelligence and Applications*, 372:2346–2353.

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.

Anish Thapaliya. 2026. egrantha.ai@chipsal 2026: Stochastic image captioning for robust hate speech detection in low-resource nepali memes. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*.

Samir Wagle, Reewaj Khanal, and Abiral Adhikari. 2026. Meme-fusion@chipsal 2026: Multimodal ablation study of hate detection and sentiment analysis on nepali memes. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*.

Bo Xu, Tingting Li, Junzhe Zheng, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. 2022. Met-meme: A multimodal meme dataset rich in metaphors. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 2887–2899.