HasNat@CHiPSAL 2026: Multimodal Hate Speech Detection in Low-Resource Nepali Memes Using Aligned Vision–Language Models

Alvee Hasan Chowdhury, Md. Abul Hasnat, Adnan Faisal
Department of Computer Science and Engineering,
Chittagong University of Engineering and Technology, Bangladesh
{u2004011, u2004005, u2004002}@student.cuet.ac.bd

Abstract

Memes, while widely used for humor and everyday expression, have increasingly been exploited to spread hate and harmful stereotypes on social media. Detecting such content has been particularly challenging because meme meaning has been conveyed jointly through images and embedded text and the problem has become harder in low-resource languages like Nepali where annotated multimodal resources have remained limited. In the CHiPSAL 2026 Shared Task on Multimodal Hate and Sentiment Understanding in Low-Resource Memes (Subtask-A), we have focused on identifying hate speech in Nepali-only memes by modeling both textual and visual cues. We have benchmarked three multimodal model families-ViT-B-32 (OpenCLIP), AltCLIP and BLIP2+mT5-under baseline, preprocessing, augmentation and combined settings. Among all systems, an AltCLIP-based late-fusion classifier has achieved the best validation performance with a macro F1-score of 0.66, indicating that aligned vision-language representations have been effective for meme-level understanding in Nepali. ViT-B-32 has provided a strong Baseline with a macro F1-score of 0.65, while BLIP2+mT5 has performed best when preprocessing and augmentation have been combined, achieving a macro F1-score of 0.62. Overall, our findings have highlighted the importance of multimodal modeling and meme-safe transformations for robust hate speech detection in low-resource meme content.

1 Introduction

Memes have been widely used on social media for humour and cultural commentary, typically combining an image with embedded text (Ramamoorthy et al., 2022). However, the same format has increasingly been exploited to spread hate and misinformation, making automated detection essential. Hate speech detection has remained challenging because it is often subtle, context-dependent and may require world knowledge (Kiela et al., 2021). While progress has been made for high-resource languages such as English, low-resource languages like Nepali have remained underexplored despite the growing presence of Nepali and code-mixed meme content (Thapa et al., 2025a).

The CHiPSAL 2026 Shared Task on Multimodal Hate and Sentiment Understanding in Low-Resource Memes has focused on Nepali memes and has required models to jointly analyze both the image and its embedded Nepali text (Thapa et al., 2026a; Sarveswaran et al., 2026). In this work, we have evaluated three multimodal model families: ViT-B-32 (OpenCLIP), AltCLIP, and BLIP2+mT5. We compared these models under baseline, preprocessing, augmentation, and combined settings. AltCLIP (Baseline) has achieved the best validation performance with an F1-score of 0.66, highlighting the effectiveness of aligned vision-language representations for Nepali meme hate-speech detection. We have also examined preprocessing and augmentation variants to study how input cleaning and synthetic diversity have influenced robustness and generalization across backbones.

The core contributions of our research work have been as follows:

- We benchmarked multiple multimodal backbones (OpenCLIP, AltCLIP, BLIP2+mT5) under consistent experimental settings for Nepali meme hate speech detection.
- We analyzed the impact of preprocessing and augmentation variants on each backbone to study robustness and generalization.
- We identified AltCLIP (Baseline) as the strongest model in our experiments, achieving the highest validation macro F1-score of 0.66.

Detailed implementation information has been made available in the GitHub repository: https://github.com/alvee-hasan/Nepali_Meme_Hate_Speech_Notebooks

Proceedings of the Second workshop on Challenges in Processing South Asian Languages (CHiPSAL2026) @ LREC, pages 237–243

2 Related Work

Hate speech detection has been widely studied and recent work has emphasized that online harm is often multimodal, where text and images jointly convey abusive meaning. Benchmarks such as the Hateful Memes Challenge have driven progress on models that must reason over both modalities (Kiela et al., 2020, 2021). In parallel, shared tasks and datasets on meme sentiment and emotion understanding (e.g., Memotion) and multimodal meme sentiment resources (e.g., MemoSen) further highlighted the importance of combining visual and textual cues for robust prediction (Sharma et al., 2020; Ramamoorthy et al., 2022). Beyond hate, multimedia abuse benchmarks such as MAMI expanded the evaluation landscape to related harms in memes, reinforcing the need for multimodal modeling approaches (Fersini et al., 2022).

For low-resource and multilingual settings, cross-lingual transfer has been explored to reduce annotation costs and improve generalization when labeled data is limited (Bigoulaeva et al., 2021). In the Nepali context, prior work has released annotated hate speech data and analyzed hate phenomena in Nepali discourse, providing foundations for building moderation systems in this language (Thapa et al., 2023). Recently, multimodal prompt- based approaches have also been investigated for code-mixed and low-resource memes, suggesting that strong pretrained multimodal representations can be adapted effectively with careful task design (Thapa et al., 2025a,b). Building on these directions, the CHiPSAL 2026 shared task focuses on multimodal hate and sentiment understanding in Nepali-only memes, encouraging systems that jointly model meme images and Nepali text for hate speech detection (Subtask A) and sentiment analysis (Subtask B) (Thapa et al., 2026a; Sarveswaran et al., 2026). We also note that the annotation schema used in this shared-task data follows prior multimodal hate resources, including CrisisHateMM (Bhandari et al., 2023).

3 Data Description

The Nepali Meme Hate Speech Detection dataset (Thapa et al., 2026a,b; Bhandari et al., 2023) consists of paired meme images and Nepali text, organized into Train, Development and Test splits. It contains 1,068 training samples, 133 development samples and 134 test samples, with 1,335 corresponding images in total. The split-wise distribution is shown in Table 1.

Label	Train	Development	Test
Hate	720	98	–
Non-Hate	348	35	–
Total	**1068**	**133**	**134**

Table 1: Label-wise Distribution of Training, Development and Test Data

3.1 Data Augmentation

Data augmentation was explored only in the AltCLIP variants that explicitly included augmentation. In the baseline AltCLIP setting, no augmentation was applied. Images were loaded, converted to RGB and passed directly to the AltCLIP processor, while text was used in its original form after basic safe handling. This preserved the original meme layout, caption readability and image-text alignment. Among all AltCLIP variants, this baseline achieved the best overall performance, with 0.66 precision, 0.66 recall, 0.70 accuracy and 0.66 macro F1, suggesting that preserving the original meme structure was most beneficial.

In the augmentation-only AltCLIP setting, augmentation was applied only to images; no textual augmentation was used. The image pipeline consisted of RandomResizedCrop, RandomHorizontalFlip, ColorJitter and RandomRotation. Training data were also balanced by oversampling each class to 1000 samples per fold. Although this variant achieved high precision (0.85), recall dropped to 0.55 and macro F1 fell to 0.49. This indicates that stronger visual transforms likely distorted meme-specific cues, especially embedded text and fine caption layout.

In the preprocessing + augmentation AltCLIP setting, augmentation remained image-only but was made more conservative. The training pipeline used slight upscaling followed by random cropping, RandomHorizontalFlip and light ColorJitter, while random rotation and stronger distortions were avoided to better preserve text readability. This setting also used balanced oversampling, with 1200 samples per class per fold. Even so, it remained below the baseline, reaching 0.62 precision, 0.60 recall, 0.68 accuracy and 0.61 macro F1.

Overall, the AltCLIP results suggest that augmentation was not beneficial for this task. Because meme classification depends heavily on caption visibility, text placement and precise image-text correspondence, even moderate visual transformations may remove or weaken important task-relevant in-

formation.

4 Methodology

4.1 Problem Formulation

The task is to classify a meme as *Hate* or *Non-Hate* using multimodal Nepali meme data(Table 1). Given a meme instance $m = (t, i)$, where t is the embedded Nepali text and i is the associated image, we learn a mapping:

$$f(t, i) \rightarrow \{0, 1\},$$

where 0 denotes Non-Hate and 1 denotes Hate. The objective is to jointly model textual and visual cues so that implicit and explicit hate expressions can be detected more reliably than with a single modality.

4.2 Data Preprocessing

For text, missing values were replaced with empty strings and all entries were converted to string format. In the baseline AltCLIP setting, preprocessing was intentionally light so that the model could preserve the original meme wording, spelling variation, punctuation and code-mixed cues. The text was then tokenized using the AltCLIP processor with a maximum sequence length of 77, applying truncation and fixed-length padding. No transliteration was applied; Nepali text was retained in its original Devanagari form.

For the AltCLIP variants that used additional preprocessing (i.e., the *Preprocessing* and *Preprocessing + Augmentation* settings), a stronger text normalization pipeline was applied before tokenization. This pipeline included Unicode normalization, normalization of Devanagari digits to standard ASCII digits, lowercasing of Latin-script characters, regex-based removal of source-page and watermark phrases, cleanup of URLs and user handles, whitespace normalization, repetition cleanup for unusually noisy OCR strings and text-length capping. These steps were designed primarily to reduce OCR noise, suppress source leakage and standardize noisy user-generated text while preserving the main semantic content of the meme.

For images, each meme was converted to RGB and resized/normalized using the AltCLIP image processor at a resolution of 224×224. In the baseline AltCLIP model, this processor-driven pipeline was used directly. In the preprocessing-based AltCLIP variants, an additional edge-cropping step was applied before resizing in order to suppress border watermarks and page-name overlays that frequently appeared near the image margins. We also verified image availability prior to training to prevent broken samples; if image loading failed, a white fallback image was used to ensure robustness. The *Preprocessing + Augmentation* AltCLIP variant inherited the same preprocessing pipeline before any training-time augmentation was applied.

To improve reproducibility and stability, we set fixed random seeds and employed a stratified split strategy during cross-validation so that class proportions remained consistent across folds. For completeness, the other explored backbones followed broadly similar goals, namely OCR-noise reduction and watermark suppression. ViT-B/32 used a comparable CLIP-style preprocessing pipeline, whereas the BLIP-2 + mT5 variants additionally used stronger text cleaning such as repeated n-gram filtering and a separate mT5-based text encoding branch. But overall, AltCLIP (Baseline) was the best-performing model.

4.3 Multimodal Model

AltCLIP was used as the multimodal backbone. The model encodes text and images separately, producing aligned text and image embeddings. These embeddings are then concatenated and passed to a trainable MLP classifier (late fusion).

Let $\mathbf{e}_t \in \mathbb{R}^d$ and $\mathbf{e}_i \in \mathbb{R}^d$ denote the text and image embeddings, respectively. The fused representation is computed as:

$$\mathbf{z} = [\mathbf{e}_i; \mathbf{e}_t] \in \mathbb{R}^{2d},$$

followed by:

$$\hat{y} = \sigma(\text{MLP}(\mathbf{z})).$$

The AltCLIP backbone was kept frozen and only the fusion classifier was trained. This design reduced memory cost and supported stable training while retaining strong pretrained cross-modal alignment.

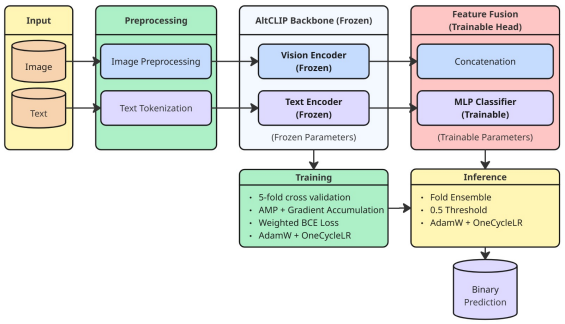

Figure 1: Overview of the multimodal framework for Nepali meme hate speech detection.

4.4 Training Setup

The model was trained on the combined Train and Development sets using 5-fold stratified cross-validation to preserve class distribution across folds. Optimization was performed using AdamW (learning rate 1×10^{-3}, weight decay 0.01) for 6 epochs per fold, with a batch size of 8 and gradient accumulation over 2 steps to maintain memory efficiency. To address class imbalance, we applied weighted binary cross-entropy with fold-specific class weights computed from the training labels. For stable and efficient learning, we used mixed precision (AMP), gradient clipping (max norm 1.0) and a OneCycle learning-rate schedule. During inference, we averaged the predicted probabilities from the best checkpoint of each fold (fold ensemble) and applied a 0.5 threshold for final label prediction.

4.5 Evaluation Metrics

The models were evaluated using macro-F1 score, precision and recall to ensure balanced performance and accurate identification of hate speech in Nepali-only memes.

5 Results and Analysis

This task evaluates multimodal models using meme text and images for Nepali hate-speech detection. While several settings achieved strong validation scores, the results indicate limited generalization, motivating more robust fusion and meme-aware augmentation.

5.1 Task: Hate Speech Detection in Nepali-only Memes

Model	Setting	P	R	F1
ViT-B-32	Preprocessing	0.62	0.60	0.61
	Augmentation	0.61	0.61	0.61
	Preprocessing + Augmentation	0.64	0.63	0.63
	Baseline	**0.65**	**0.64**	**0.65**
AltCLIP	Preprocessing	0.63	0.62	0.62
	Augmentation	**0.85**	0.55	0.49
	Preprocessing + Augmentation	0.62	0.60	0.61
	Baseline	0.66	**0.66**	**0.66**
BLIP2+mT5	Preprocessing	0.60	0.57	0.57
	Augmentation	0.59	0.58	0.58
	Preprocessing + Augmentation	**0.63**	**0.62**	**0.62**
	Baseline	0.61	0.61	0.61

Table 2: Performance comparison across model families and settings. Best results per model family are highlighted in bold.

Table 2 reports the performance of three multimodal backbones: ViT-B-32, AltCLIP, and BLIP2+mT5. The models are evaluated under four data settings: baseline, preprocessing, augmentation, and preprocessing with augmentation. For the ViT-B-32 family, the baseline configuration yields the strongest results with P=0.65, R=0.64 and F1=0.65, indicating that the default training setup already provides a robust CLIP-based baseline for Nepali meme hate-speech detection. In the AltCLIP experiments, the baseline achieves the best overall performance within the family (P=0.66, R=0.66, F1=0.66), demonstrating improved alignment of visual and textual representations compared to ViT-B-32. Although the augmentation-only AltCLIP setting achieved the highest precision (0.85), its substantially lower recall (0.55) reduced the overall F1-score, indicating that many positive memes were still missed. This was likely caused by the combination of residual class imbalance and an aggressive augmentation setup, including random resized cropping, horizontal flipping, color jitter and rotation, together with a very low learning rate, which may have weakened fine-grained meme text cues and reduced sensitivity. These transforms likely made the model more conservative by weakening subtle meme-specific cues, especially embedded text and local caption structure. As a result, the model tended to detect only the clearest hateful cases, which increased precision but reduced recall. A milder setup with gentler cropping, little or no flipping or rotation, lighter color jitter and a slightly higher learning rate would likely improve recall. For BLIP2+mT5, the preprocessing + augmentation configuration performs best (P=0.63, R=0.62, F1=0.62), showing that combining input cleaning with augmentation is beneficial for this architecture. Overall, AltCLIP provides the most balanced performance among the compared families, with the baseline achieving the best results, while augmentation-only causes the largest drop. Although preprocessing + augmentation recovers some performance, it remains below the unaugmented baseline.

5.2 Parameter Setting

Table 3 has summarized the main hyperparameters used in our experiments. For ViT-B-32, we report the learning rates for the two training stages (Stage-1 linear probing and Stage-2 light fine-tuning) in separate rows to keep the table compact. AltCLIP has been trained in a head-only setting with

the backbone frozen for memory efficiency. For BLIP2+mT5, we used a two-step pipeline and reported separate batch sizes for feature extraction and classifier training.

Model	Learning Rate	Optimizer	Batch Size
ViT-B-32 – Stage 1	3×10^{-3}	AdamW	32
ViT-B-32 – Stage 2	1×10^{-5}	AdamW	32
AltCLIP (head-only)	1×10^{-3}	AdamW	8
BLIP2+mT5 (extract)	3×10^{-4}	AdamW	4
BLIP2+mT5 (classifier)	3×10^{-4}	AdamW	64

Table 3: Key hyperparameters for the three model families. ViT-B-32 uses a two-stage training strategy with separate learning rates.

6 Conclusion

In this work, we addressed CHiPSAL 2026 Subtask-A on hate speech detection in Nepali-only memes by evaluating three multimodal model families: ViT-B-32 (OpenCLIP), AltCLIP, and BLIP2+mT5. We compared these models under baseline, preprocessing, augmentation, and combined settings. Among all systems, the AltCLIP baseline achieved the best validation performance, confirming the effectiveness of strongly aligned vision-language embeddings for low-resource meme moderation. We further observed that data-centric strategies were highly model-dependent: preprocessing and augmentation sometimes reduced performance when they removed dataset-specific cues (e.g., watermarks) or distorted embedded text, leading to recall drops. In particular, AltCLIP performed best without image augmentation, while aggressive visual transforms hurt meme classification by weakening embedded text and local caption layout; milder combined augmentation reduced this damage but still underperformed the baseline. Overall, our results highlight the importance of multimodal representations and meme-safe transformations for reliable Nepali hate-speech detection. Partial unfreezing of the top AltCLIP layers may improve macro F1 by enabling limited task-specific adaptation, although the small dataset size makes overfitting a significant risk. Future work will therefore examine threshold calibration, controlled ablations, and carefully limited backbone fine-tuning to improve robustness and generalization for practical content moderation support.

Error Analysis

Although AltCLIP (Baseline) performs best overall, the augmentation-based results show that simple count balancing did not fully solve the data skew problem. Table 4 summarizes how class balancing differed across the augmentation-based settings and helps explain why oversampling did not consistently improve generalization. ViT-B/32 augmentation variants remained imbalanced because no resampling was applied, while AltCLIP and BLIP-2 balanced only the training fold or train set through oversampling, mainly inflating the minority Non-Hate class without increasing unique example diversity. As a result, the experiments remained effectively imbalanced at evaluation time, since validation still followed the original class prior. This is especially problematic for meme classification, where embedded text, caption placement, and image-text alignment carry much of the signal, so aggressive transforms can easily damage critical cues. Even beyond this, AltCLIP still makes errors on memes containing implicit or context-dependent hate, such as sarcasm, idioms, or cultural cues that go beyond literal wording. It also fails when image-embedded text is small, stylized, or partially hidden, leading to missed cues and false negatives-an issue that worsens with stronger transforms like cropping or blur that reduce readability. Moreover, score differences across preprocessing and augmentation settings are partly affected by non-identical training conditions, including batch size, learning rate, and sampling strategy, while class balancing can also alter calibration; therefore, a fixed 0.5 threshold may misclassify borderline cases and hurt macro-F1.

Limitations

Despite strong performance from the Baseline AltCLIP system, limitations remain. The dataset is imbalanced and contains source-specific artifacts (e.g., watermarks), which may encourage shortcut learning; removing them can also discard label-correlated cues. The backbone was kept mostly frozen, limiting adaptation to distribution shifts introduced by preprocessing or augmentation. Cultural context and implicit intent in memes are also difficult to capture with simple embedding fusion and thresholds were not tuned separately for each setting, so calibration shifts may affect macro-F1. For BLIP2+mT5, hardware limitations on the Tesla P100 imposed a major constraint. End-to-end BLIP-2 fine-tuning was infeasible because the

Models	Training Scope	Label 0 (Non-Hate)	Label 1 (Hate)	Reason of augmentation failure
ViT-B/32 Augmentation	combined train+val for 5-fold CV	383 → 383	818 → 818	No resampling, still imbalanced
ViT-B/32 Preprocessing + Augmentation	combined train+val for 5-fold CV	383 → 383	818 → 818	No resampling, still imbalanced
AltCLIP Augmentation	per CV training fold	306–307 → 1000	654–655 → 1000	Fold-only oversampling balance
AltCLIP Preprocessing + Augmentation	per CV training fold	306–307 → 1200	654–655 → 1200	Fold-only oversampling balance
BLIP-2 Augmentation	train only	348 → 1200	720 → 1200	Train-only oversampling balance
BLIP-2 Preprocessing + Augmentation	train only	348 → 1200	720 → 1200	Train-only oversampling balance

Table 4: Class-balancing behavior in augmentation-based settings.

full model exceeded memory limits, while 8-bit quantization could not be used due to bitsandbytes incompatibility with the P100 architecture. As a result, BLIP-2 was used only as a frozen feature extractor with a lightweight classifier, which reduced memory usage but likely capped task-specific adaptation and overall performance.

Ethical Statement

This work followed established ethical guidelines for handling sensitive hate-speech content. It aimed to improve automated hate-speech detection while respecting user privacy and fundamental rights. We acknowledge the possibility of dataset bias and label subjectivity and we have reported the limitations of our study transparently. The proposed model is intended to support human moderation rather than replace human judgment.

Acknowledgement

We have sincerely thanked the organizers of the CHiPSAL 2026 Shared Task for providing the benchmark dataset and evaluation framework. We have acknowledged (Thapa et al., 2025a) for releasing the NeMeme dataset, which forms the foundation of this study. We have expressed our sincere gratitude to the open-source community for developing and openly sharing the tools, libraries and pretrained models that have made this work possible.

References

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.

Irina Bigoulaeva, Viktor Hangya, and Alexander Fraser. 2021. Cross-lingual transfer learning for hate speech detection. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 15–25, Kyiv. Association for Computational Linguistics.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Casey A. Fitzpatrick, Peter Bull, Greg Lipstein, Tony Nelli, Ron Zhu, Niklas Muennighoff, Riza Velioglu, Jewgeni Rose, Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, Helen Yannakoudakis, Vlad Sandulescu, Umut Ozertem, Patrick Pantel, Lucia Specia, and Devi Parikh. 2021. The hateful memes challenge: Competition report. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pages 344–360. PMLR.

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 *Advances in Neural Information Processing Systems 33 (NeurIPS 2020), December 6–12, 2020, virtual*.

Sathyanarayanan Ramamoorthy, Nethra Gunti, Shreyash Mishra, Suryavardan S, Aishwarya N. Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit P. Sheth, Asif Ekbal, and Chaitanya Ahuja. 2022. Memotion 2: Dataset on sentiment and emotion analysis of memes (short paper). In *Proceedings of the Workshop on Multi-Modal Fake News and Hate-Speech Detection (DE-FACTIFY 2022), co-located with AAAI 2022*, volume 3199 of *CEUR Workshop Proceedings*.

Kengatharaiyer Sarveswaran, Surendrabikram Thapa, Ashwini Vaidya, Tafseer Ahmed Khan, and Bal Kr-

ishna 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)*.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 task 8: Memotion analysis- the visuo-lingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.

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 - 26th European Conference on Artificial Intelligence*, volume 372 of *Frontiers in Artificial Intelligence and Applications*, pages 2346–2353. IOS Press.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Kritesh Rauniyar, Laxmi Thapa, Surabhi Adhikari, Kristina T. Johnson, Kengatharaiyer Sarveswaran, Bal Krishna Bal, and Usman Naseem. 2026a. 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 (CHiPSAL)*.

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.

Surendrabikram Thapa et al. 2026b. therealthapa/chipsal26-memes: Shared task on multimodal hate and sentiment understanding in low-resource memes (chipsal 2026). GitHub repository. Accessed: 2026-02-27.