Unigoa@CHiPSAL 2026: Early vs Late Fusion for Multimodal Hate and Sentiment Detection in Nepali Memes

Ashweta A. Fondekar, Milind M. Shivolkar, Jyoti D. Pawar
Parvatibai Chowgule College of Arts and Science(Autonomous); Goa Business School, Goa University
aaf016@chowgules.ac.in, {dcst.ashweta, milind.shivolkar, jdp}@unigoa.ac.in

Abstract

Internet memes pose significant challenges for automatic content moderation due to the interaction of visual and textual cues, sarcasm, and cultural context. In this work, we participate in the CHiPSAL 2026 shared task on multimodal hate and sentiment understanding in Nepali memes. The task consists of two subtasks: binary hate speech detection and three-class sentiment classification. We investigate both early-fusion and late-fusion multimodal architectures. Our primary system employs a late-fusion dual-encoder architecture combining XLM-RoBERTa for multilingual text representation and CLIP for visual encoding. We further evaluate an early-fusion ViLT-based joint vision–language transformer using NepBERTa tokenization as a baseline. Experimental results show that late-fusion models consistently outperform early-fusion architectures, particularly for code-mixed memes containing Devanagari Nepali and Roman-script English text. Our best system achieves a Macro-F1 of 0.6564 for hate speech detection and 0.4859 for sentiment classification. We provide analysis highlighting the challenges of multilingual code-mixing, sarcasm, and implicit sentiment in low-resource multimodal settings.

Keywords: Multimodal Learning, Multimodal Fusion, Hate Speech Detection, Sentiment Analysis, Code-Mixed NLP, Nepali Memes

1. Introduction

Internet memes have emerged as a dominant form of communication on social media platforms. By combining images, short text, humor, and cultural references, memes enable users to express opinions and emotions in a compact and highly shareable format. However, the same characteristics that make memes engaging also make them difficult to moderate automatically. The meaning of a meme often depends on the interaction between visual content and text, and the intended message may be conveyed through sarcasm, irony, or implicit cultural context. Consequently, detecting hate speech or identifying sentiment in memes is considerably more challenging than in traditional text-only content (Parihar et al., 2021).

Despite recent progress in multimodal learning, most prior research has focused on English-language datasets. Low-resource languages such as Nepali remain underrepresented in multimodal content moderation research. Nepali social media discourse presents additional challenges due to informal language usage, frequent code-mixing between Nepali (Devanagari script) and English (Latin script), and culturally grounded humor. These characteristics significantly increase the difficulty of automatically identifying harmful or sentiment-bearing content.

This work participates in the CHiPSAL 2026 shared task on multimodal hate and sentiment understanding in Nepali memes (Thapa et al., 2026; Sarveswaran et al., 2026). The dataset builds upon previous research on multimodal meme understanding and code-mixed content (Thapa et al., 2025a,b) and follows annotation schemes inspired by CrisisHateMM (Bhandari et al., 2023). The shared task includes two subtasks:

- **Subtask A:** Binary hate speech detection.
- **Subtask B:** Three-class sentiment classification (positive, neutral, negative).

We investigate both early-fusion and late-fusion multimodal architectures. Our primary system uses a late-fusion dual-encoder architecture combining XLM-RoBERTa with CLIP. We additionally evaluate an early-fusion ViLT-based joint vision–language transformer using NepBERTa tokenization as a baseline. This design enables us to study the role of multilingual pretraining and fusion strategies in low-resource multimodal settings.

Our contributions are threefold:

- We propose a late-fusion multimodal architecture combining XLM-RoBERTa and CLIP for hate speech and sentiment detection in Nepali memes.
- We introduce a stabilised training strategy for sentiment classification using weighted sampling, soft focal loss, gradient accumulation, and discriminative learning rates.
- We provide an analysis of multilingual code-mixing and multimodal challenges in Nepali meme understanding.

Proceedings of the Second workshop on Challenges in Processing South Asian Languages (CHiPSAL2026) @ LREC, pages 229–236

Paper Structure. The remainder of this paper is organised as follows. Section 2 reviews related work on multimodal hate speech and sentiment analysis. Section 3 describes the dataset and shared task. Section 4 presents the proposed multimodal architectures and training strategies. Section 5 details the experimental setup. Section 6 reports the results and analysis, and Section 7 concludes the paper.

2. Related Work

Early research on hate speech detection primarily focused on text-only analysis, highlighting linguistic challenges and societal implications (Parihar et al., 2021). With the rapid growth of image-centric social media, research has increasingly shifted toward multimodal analysis of text embedded in images.

The CrisisHateMM dataset introduced large-scale annotations for multimodal hate speech and demonstrated the importance of combining visual and textual signals (Bhandari et al., 2023). The Hateful Memes benchmark further emphasized the need for joint vision–language reasoning in meme understanding (Kiela et al., 2020). More recent work has explored multimodal meme understanding in code-mixed and low-resource settings using prompt-based and retrieval-augmented approaches (Thapa et al., 2025a,b).

Advances in multimodal representation learning have been driven by vision–language pretraining models such as CLIP (Radford et al., 2021) and ViLT (Kim et al., 2021), which learn shared representations across image and text modalities. Multimodal sentiment analysis has also been widely studied (Zadeh et al., 2018). In parallel, multilingual transformer models such as XLM-R (Conneau et al., 2020) have significantly improved modelling of code-mixed and low-resource languages.

The CHiPSAL shared task extends this line of research by focusing specifically on Nepali memes and low-resource multimodal moderation (Thapa et al., 2026; Sarveswaran et al., 2026). Our work builds upon these efforts by comparing early- and late-fusion multimodal architectures for Nepali meme understanding.

3. Dataset and Task Description

3.1. NeMeme Dataset Overview

We conduct experiments on the *NeMeme* dataset (Thapa et al., 2025a), a benchmark for multimodal meme understanding in Nepali. The dataset contains memes collected from social media platforms including Twitter, Reddit, Instagram, Facebook, and Threads. Each sample consists of an image paired with OCR-extracted text, enabling joint multimodal analysis of visual and textual information.

The memes are annotated by trained annotators for both hate speech and sentiment, and the dataset reports substantial inter-annotator agreement, indicating reliable labeling quality. The content reflects real-world Nepali social media discourse, including humor, sarcasm, informal language usage, and culturally grounded references.

Although the dataset is described as Nepali-only, manual inspection reveals frequent code-mixing between Nepali written in Devanagari script and English words written in Roman script.

3.2. Task Definition

The shared task consists of two classification problems:

- **Subtask A – Hate Speech Detection:** A binary classification task where memes are labeled as *hateful* or *non-hateful*.
- **Subtask B – Sentiment Classification:** A three-class classification task where memes are labeled as *negative*, *neutral*, or *positive*.

3.3. Linguistic Characteristics of the Dataset

Although the task is described as Nepali-only, the dataset reflects authentic Nepali social media usage, where English lexical insertions frequently appear within Nepali sentences. Manual inspection reveals substantial code-mixing between Nepali written in Devanagari script and English words written in Roman script.

Many memes contain English sentiment words, sarcasm markers, informal expressions, and references to meme pages embedded within Nepali discourse. In several cases, sentiment-bearing expressions appear in English, while the surrounding syntactic structure remains Nepali.

Such code-mixing mirrors real-world digital communication practices in Nepal and introduces additional challenges for monolingual models. The presence of mixed scripts, informal spelling variations, and culturally grounded humor increases the semantic complexity of both hate speech detection and sentiment classification. These linguistic characteristics strongly motivate the use of multilingual and multimodal encoders in our experiments.

3.4. Dataset Splits and Class Distribution

Tables 1 and 3 summarize the data splits for both subtasks, while Tables 2 and 4 present the label distributions in the training sets. The hate speech task exhibits noticeable class imbalance, which motivated the use of threshold tuning. For sentiment

Split (Hate Task)	Samples
Train	1068
Validation	133
Test	134

Table 1: Data split for hate speech detection.

Label (Hate Task)	Train Count
Hate (1)	720
Non-hate (0)	348

Table 2: Training label distribution for hate speech detection.

classification, class-weighted loss was applied to address imbalance across the three classes.

The validation sets follow a similar distribution. For hate speech detection, the validation set contains 98 hateful and 35 non-hateful memes. For sentiment classification, the validation set contains 39 negative, 65 neutral, and 29 positive memes.

4. Methodology

To address the two subtasks, we design multimodal systems that jointly model visual and textual information. While Subtask A focuses on binary hate detection, Subtask B involves three-way sentiment classification and presents additional challenges due to ambiguity, sarcasm, and code-mixed Nepali–English text. We therefore explore both early-fusion and late-fusion multimodal architectures to analyze the impact of fusion strategies and multilingual text encoders in low-resource meme understanding.

4.1. Subtask A: Multimodal Hate Speech Detection

We model hate speech detection as a binary multimodal classification problem. Each meme consists of an image and associated OCR text. Our primary system follows a late-fusion design with separate encoders for text and images.

Text is encoded using **XLM-RoBERTa-Base**, a multilingual transformer pretrained on large multilingual corpora. Token embeddings are aggregated using masked mean pooling to obtain a fixed-length sentence representation.

Images are encoded using **CLIP ViT-B/32**, which

Split (Sentiment Task)	Samples
Train	1061
Validation	133
Test	133

Table 3: Data split for sentiment classification.

Label (Sentiment Task)	Train Count
Negative (0)	341
Neutral (1)	473
Positive (2)	247

Table 4: Training label distribution for sentiment classification.

produces a pooled visual embedding. The text and image embeddings are concatenated and passed to a two-layer MLP classifier to produce binary logits.

Because the dataset is imbalanced, we apply **validation-based threshold tuning**. Instead of using the default decision threshold of 0.5, we select the probability threshold that maximizes Macro-F1 on the validation set before generating test predictions.

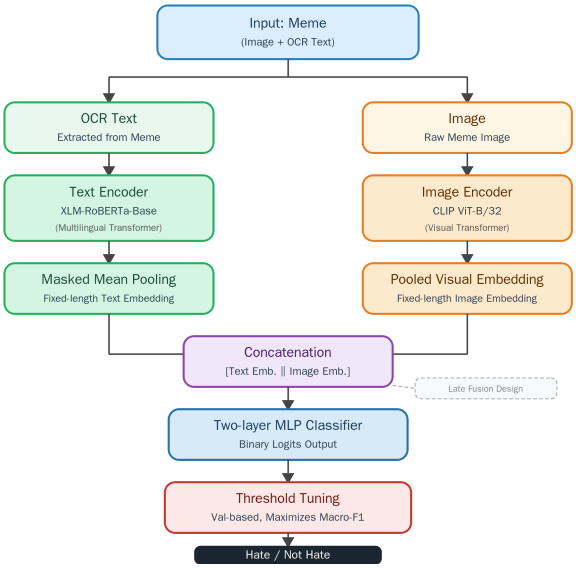

Figure 1: Late-fusion architecture for Subtask A (hate speech detection).

4.2. Subtask B: Multimodal Sentiment Classification

Sentiment classification is formulated as a three-class problem (negative, neutral, positive). We adopt a stronger variant of the fusion architecture using **XLM-RoBERTa-Large** for text encoding while retaining CLIP for image encoding.

Due to increased task difficulty and severe class imbalance, several training stabilisation techniques are introduced:

- WeightedRandomSampler for balanced mini-batches
- Soft Focal Loss with class-dependent weighting

- Gradient accumulation to increase effective batch size
- Discriminative learning rates for encoder vs classifier
- Freezing encoders during early training epochs

The fused multimodal embedding is passed to an MLP classifier with LayerNorm and GELU activations.

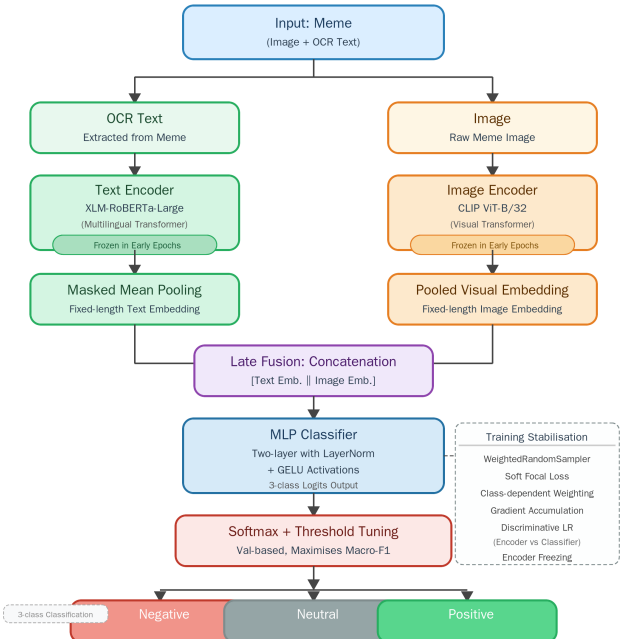

Figure 2: Late-fusion architecture for Subtask B (sentiment classification).

4.3. ViLT + NepBERTa (Early-Fusion Baseline)

To provide a comparative baseline, we evaluate a joint vision–language transformer using the **ViLT** architecture with **NepBERTa tokenization**. In this early-fusion approach, image patches and text tokens are processed jointly within a single transformer backbone, enabling cross-modal interactions at multiple layers.

The token embedding layer is resized to match the NepBERTa tokenizer vocabulary. The model is fine-tuned with task-specific classification heads for both subtasks using cross-entropy loss. This model serves as an early-fusion baseline against which our late-fusion XLM-R + CLIP models are compared. NepBERTa is selected for this model because it is specifically pretrained on Nepali text, allowing us to evaluate whether language-specific tokenization benefits early-fusion multimodal learning.

4.4. Model Comparison Strategy

We compare two multimodal fusion paradigms:

- **Early fusion:** ViLT + NepBERTa joint vision–language transformer.
- **Late fusion:** XLM-RoBERTa + CLIP dual-encoder architecture.

This comparison allows us to analyze how multimodal interaction strategies and multilingual pretraining influence performance on Nepali-dominant code-mixed memes. In particular, we investigate whether stronger multilingual text representations combined with late fusion can better capture sarcasm, humor, and culturally grounded expressions commonly found in Nepali memes. We intentionally restrict the comparison to two representative architectures to isolate the impact of multimodal fusion strategy under limited shared-task data. The ViLT + NepBERTa model represents a fully joint vision–language transformer trained in an early-fusion manner, while the XLM-R + CLIP architecture represents a dual-encoder late-fusion paradigm leveraging strong unimodal pretraining. This focused comparison enables clearer analysis of fusion strategy effectiveness without confounding factors introduced by multiple model variants.

5. Experimental Setup

5.1. Implementation Details

All models are implemented using PyTorch and the HuggingFace Transformers framework. Images are processed using the CLIP image processor, which performs resizing and normalization before being passed to the CLIP ViT-B/32 encoder. Meme text is tokenized using the corresponding XLM-RoBERTa tokenizers. Training is performed on a single GPU using mixed-precision (Automatic Mixed Precision, AMP) to improve computational efficiency and reduce memory usage.

5.2. Training Configuration

Subtask A: Hate Speech Detection. The XLM-RoBERTa-Base + CLIP late-fusion model is trained using the AdamW optimizer with a cosine learning-rate scheduler and linear warm-up. The learning rate is set to 2×10^{-5} with weight decay of 0.01.

To stabilize training, both text and image encoders are frozen during the first two epochs, allowing the classifier head to learn task-specific representations before full joint fine-tuning. Training employs automatic mixed precision and gradient clipping. Early stopping is applied based on validation Macro-F1. To address class imbalance, we perform validation-based threshold tuning and select

Component	Configuration
Text encoder (Subtask A)	XLM-R Base
Text encoder (Subtask B)	XLM-R Large
Image encoder	CLIP ViT-B/32
Optimizer	AdamW
Learning rate (head)	2×10^{-5}
Learning rate (encoders, B)	1×10^{-6}
Weight decay	0.01
Scheduler	Cosine (A), Linear (B) + warm-up
Batch size	16 (A), 8 (B)
Gradient accumulation	4 steps (Subtask B)
Epochs	10 (A), 20 (B)
Freeze epochs	2 (A), 5 (B)
Mixed precision	Automatic Mixed Precision (AMP)
Imbalance handling (A)	Validation threshold tuning
Imbalance handling (B)	Weighted sampler + Soft Focal Loss
Gradient clipping	1.0

Table 5: Training configuration of the proposed multimodal systems.

the probability threshold that maximizes Macro-F1 for final test prediction.

Subtask B: Sentiment Classification. For sentiment classification, we adopt a stronger XLM-RoBERTa-Large + CLIP fusion model due to the increased difficulty of three-class sentiment prediction. Training incorporates several stabilisation techniques designed for imbalanced and low-resource multimodal learning.

Balanced minibatches are generated using a WeightedRandomSampler, and Soft Focal Loss with class-dependent weighting is applied to improve minority-class recall without sacrificing precision. Encoders are frozen for the first five epochs and subsequently unfrozen for joint fine-tuning using discriminative learning rates for the classifier head and encoders. Gradient accumulation is applied to increase the effective batch size while maintaining GPU memory constraints. Early stopping is again based on validation Macro-F1.

Overall, the training strategy is designed to stabilize multimodal optimisation, mitigate class imbalance, and ensure fair comparison between early- and late-fusion architectures under limited shared-task data.

Due to the limited size of the shared-task dataset, we did not perform a full ablation study of individual training strategies. The stabilisation techniques described above follow widely adopted best practices for imbalanced multimodal classification and were applied consistently across experiments to ensure stable and fair comparison between fusion architectures.

6. Results and Analysis

6.1. Main Results

Table 6 summarizes the performance of the proposed systems on the shared task test set.

The results clearly show that the **XLM-RoBERTa + CLIP late-fusion model** outperforms the **ViLT + NepBERTa early-fusion baseline** on both subtasks. For Subtask A, the fusion model achieves a Macro-F1 of **0.6564**, improving over the ViLT baseline by more than 3.6 points. For Subtask B, the improvement is larger, with the fusion model achieving a Macro-F1 of **0.4859** compared to 0.4042 for the baseline.

6.2. Leaderboard Submission

We submitted our best-performing late-fusion systems to the official CHiPSAL 2026 evaluation server. Our system obtained the following official rankings:

- **Subtask A (Hate Speech Detection): Rank 7**
- **Subtask B (Sentiment Classification): Rank 8**

These rankings indicate that the proposed late-fusion approach provides a competitive and stable baseline under shared-task conditions.

6.3. Subtask Comparison

Hate speech detection consistently outperforms sentiment classification across all evaluated models. This difference reflects the intrinsic complexity of the tasks. Hateful memes often contain relatively explicit lexical or visual indicators (e.g., offensive wording or stereotypical imagery), enabling clearer decision boundaries. In contrast, sentiment in memes is frequently conveyed through humor, irony, exaggeration, or culturally grounded references, making polarity detection inherently more ambiguous. The performance gap between Subtask A and Subtask B therefore highlights the increased semantic complexity of multimodal sentiment understanding.

6.4. Early vs Late Fusion

The comparison between ViLT (early fusion) and XLM-R + CLIP (late fusion) reveals a consistent advantage for late-fusion modeling in this dataset setting.

The early-fusion ViLT architecture processes image and text jointly within a single vision–language transformer backbone, enabling rich cross-modal interaction. However, such joint transformers typically require substantial task-specific data to effectively learn cross-modal alignment. Under limited shared-task data, early-fusion models may struggle to fully exploit multimodal interactions.

To better support Nepali text, we replace the default tokenizer with the NepBERTa tokenizer and resize the ViLT embedding layer accordingly. The

Model	Subtask A				Subtask B			
	Acc	Pre	F1	Rec	Acc	Pre	F1	Rec
XLM-RoBERTa + CLIP	**0.7164**	0.6741	**0.6564**	0.6495	**0.4887**	0.4858	**0.4859**	0.5037
NepBERTa + ViLT	0.6493	0.6186	0.6203	0.6285	0.4286	0.4053	0.4042	0.406

Table 6: Performance metrics across Subtask A (Hate Speech Detection) and Subtask B (Sentiment Classification).

transformer encoder itself remains the original ViLT backbone.

In contrast, the late-fusion approach leverages two independently pretrained encoders:

- XLM-RoBERTa for multilingual text representation
- CLIP for robust visual feature extraction

By combining strong unimodal representations learned from large-scale pretraining, the late-fusion model benefits from stable features before multimodal classification. Our findings therefore suggest that late-fusion architectures are particularly effective under low-resource multimodal conditions, rather than implying a universal superiority over early fusion.

6.5. Impact of Multilingual Text Modeling

The strong performance of XLM-R based models highlights the importance of multilingual pretraining for Nepali meme understanding. Qualitative inspection reveals frequent English–Nepali code-mixing and script mixing between Devanagari Nepali and Roman-script English. Many memes include English lexical insertions (e.g., “Free WiFi”, “best friend”, page tags such as “SARCASM” or “TROLL”) embedded within Nepali sentences. These mixed-script patterns can be challenging for monolingual models and help explain why multilingual encoders capable of handling cross-script tokenization provide consistent advantages.

6.6. Error Analysis

Despite the improvements achieved by the late-fusion model, several recurring error patterns were observed:

- **Sarcasm and irony:** Memes explicitly marked with page-level cues such as “SARCASM” or “TROLL” are sometimes misclassified when the literal OCR text appears neutral or positive.
- **Implicit sentiment and cultural context:** Certain memes rely heavily on shared cultural knowledge or visual humor, making polarity difficult to infer from text alone.
- **Code-mixed and noisy tokens:** Frequent mixing of Devanagari Nepali with Roman-script English, along with URLs, hashtags, and page tags, introduces lexical variability that weakens token-level sentiment cues.

6.7. Discussion

Overall, the findings underscore the importance of multilingual pretraining and strong unimodal representations for multimodal meme understanding in low-resource settings. While hate speech detection benefits from more explicit lexical and visual cues, sentiment classification requires deeper multimodal reasoning and improved cross-modal alignment. These results suggest that future work should explore larger multilingual multimodal pretraining and more data-efficient early-fusion architectures for low-resource meme understanding.

7. Conclusion

In this paper, we presented our systems for the CHiPSAL 2026 shared task on multimodal hate and sentiment understanding in Nepali memes. We conducted a focused comparison between early-fusion (ViLT + NepBERTa) and late-fusion (XLM-RoBERTa + CLIP) multimodal architectures and demonstrated that the late-fusion dual-encoder model consistently achieves superior performance across both subtasks.

Our results highlight the importance of multilingual pretraining for handling real-world Nepali meme content, which frequently exhibits script-level code-mixing between Devanagari Nepali and Roman-script English. While hate speech detection achieved moderate performance, sentiment classification remains considerably more challenging due to sarcasm, humor, implicit emotional cues, and culturally grounded references.

Overall, our findings indicate that multilingual dual-encoder architectures provide a strong and stable baseline under limited shared-task data conditions. Future work will explore improved cross-modal attention mechanisms, dedicated code-mixed tokenization strategies, larger multilingual multimodal pretraining, and data augmentation techniques to further enhance performance in low-resource meme understanding.

Ethical Considerations

This work focuses on multimodal content moderation in Nepali memes, which may contain offensive, hateful, or sensitive material. The dataset used in this shared task was released by the organizers for research purposes, and we strictly followed the data usage guidelines provided with the shared task. No additional data collection or annotation involving human participants was performed by the authors.

We acknowledge that automated hate speech detection systems can introduce biases and may incorrectly label benign or culturally specific expressions as harmful. Such errors may disproportionately affect certain communities or linguistic groups. Therefore, the proposed models are intended solely for research and benchmarking purposes and should not be deployed in real-world moderation systems without thorough human oversight.

Additionally, the presence of script-level code-mixing (Devanagari Nepali with Roman-script English) introduces variability that may not be fully captured by current tokenization strategies. Future work should explore dedicated code-mixed tokenizers and multilingual multimodal pretraining for South Asian languages.

The goal of this work is to support safer online spaces and to advance research in low-resource languages, while recognizing the importance of fairness, transparency, and responsible deployment of automated moderation technologies.

Limitations

This work has several limitations. First, the dataset size is relatively small compared to large-scale multimodal datasets, which restricts the ability of deep models to fully generalize. Second, the models rely on pretrained multilingual encoders that may not fully capture cultural nuances, regional humor, or evolving internet slang in Nepali memes.

Third, sentiment classification results remain modest, highlighting the difficulty of modeling sarcasm, irony, and implicit emotional cues in memes. Many memes convey meaning through subtle visual symbolism or cultural context that current models cannot fully interpret.

Finally, the experiments focus on benchmark performance within a shared task setting. Additional work is required to evaluate robustness, fairness, and cross-domain generalization before real-world deployment.

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.

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

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.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR.

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.

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.

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)*.

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 (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.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.