Cuet Yet Another Baseline@CHiPSAL LREC 2026: Shared Task on Multimodal Sentiment Understanding in Low-Resource Memes

Rotna Dipika Debnath, Shahrin Afroz Hoque Ruhi, Ayesha Labiba, Arpita Mallik, Hasan Murad
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u2104002, u2104015, u2104008, u2004023}@student.cuet.ac.bd
hasanmurad@cuet.ac.bd

Abstract

Memes serve as a method to express feelings such as humor, sarcasm, and diverse viewpoints. The task of identifying sentiment in memes is becoming increasingly complex, particularly in low-resource languages like Nepali where memes often combine images, texts, and code-mixed language. However, multimodal methods for sentiment analysis in Nepali memes seem to be insufficient. In this paper, we present our system for the Subtask B(Sentiment Analysis) for Shared Task on Multimodal Hate and Sentiment Understanding in Low-Resource Memes@CHiPSAL LREC 2026. We implement various unimodal models, such as XLM-RoBERTa-large,MuRIL-base, Twitter-XLM-R for text. Moreover, we incorporate BLIP-2 captions to enhance visual-text understanding and adopted a multimodal approach that fuses textual embeddings, image embeddings, caption embeddings, and similarity scores. The fused features process through cross-attention and a dense neural network for classification, with focal loss and class weighting used to improve performance. Our approach achieved a macro F1 score of 0.50 securing 7th place and highlighting the importance of cross-modal interaction and large-scale pretrained vision-language models for robust meme understanding in sentiment analysis.

Keywords: Multimodal Sentiment Analysis, Low-Resource Languages, Nepali NLP, Vision–Language Models, Cross-Modal Attention,BLIP-2,CLIP,XLM-RoBERTa,Code-Mixed Text,Multimodal Fusion

1. Introduction

The concept of a "meme" was first introduced to describe genetic traits that are shaped by the sociocultural environment. Originally an academic term, this phrase was later adopted by users on the internet. A look at Google Trends indicates a significant increase in interest in the topic since 2011, with over 1.9 million search results for 'Internet meme', indicating their considerable cultural relevance (Shifman, 2013).

Memes are often shared with humorous or satirical intent and have emerged as a popular medium for expressing opinions and ideas. Analyzing memes is crucial as they convey sentiments and offers insights into public opinion, trends, and social dynamics (Nguyen et al., 2022). However, their growing popularity has also led to the spread of offensive and hateful content, posing challenges for content moderation (Parihar et al., 2021). Research has increasingly focused on analysis that considers both textual and visual elements. Nevertheless, most existing studies emphasize high-resource languages with limited work on low-resource languages such as Nepali. Despite the strong presence of politically and socially driven memes in Nepali online spaces, research on Nepali and code-mixed memes remains scarce, highlighting the need for more robust multimodal approaches in this context. Figure 1 shows examples of online memes that contain both text and image.

The purpose of this paper has been to perform sentiment analysis on Nepali memes. The CHiPSAL@ACL 2026 conference (Thapa et al., 2026) has introduced a dataset under Subtask B (Sentiment Analysis) of the Shared Task on Multimodal Hate and Sentiment Understanding in Low-Resource Memes(Sarveswaran et al., 2026) which has consisted of multimodal memes annotated with sentiment labels as negative, neutral, and positive where each instance includes both textual content and an associated image, enabling the investigation of multimodal fusion strategies for sentiment understanding in a low-resource language setting.

To achieve our objective, we have designed a multimodal pipeline combining OCR-extracted text, BLIP-2 captions, CLIP ViT-L/14 image–text embeddings and XLM-RoBERTa-large for multilingual text. A cross-attention mechanism has enabled textual features to become image-aware, while focal loss with class weighting, ensemble learning, and test-time augmentation have improved robustness. Experimental results has shown that the proposed approach effectively captures subtle cross-modal cues to determine sentiment in Nepali memes. With an F1 score of 0.50, our approach has ranked 7th in the competition. The main contributions of this work have been:

- We have integrated XLM-RoBERTa, BLIP-2, and CLIP with cross-attention-based fusion in

Proceedings of the Second workshop on Challenges in Processing South Asian Languages (CHiPSAL2026) @ LREC, pages 292–299

(a) Positive

(b) Negative

(c) Neutral

Figure 1: Examples of memes representing different sentiment classes.

a multimodal architecture.

- We have created a preprocessing pipeline that combines similarity-aware visual features, generated captions, and OCR-extracted text.
- For robust evaluation, we have used stratified 4-fold cross-validation with test-time augmentation and ensemble weighting.
- We have systematically investigated OCR noise filtering strategies for low-resource Nepali meme text, demonstrating that BLIP-2 visual captioning provides inherent robustness to OCR noise, making explicit filtering unnecessary.

2. Related Work

Several studies have addressed the difficulties of identifying sentiment and harmful content embedded within memes in the current research on multimodal meme understanding. Prior studies in this domain can be categorized based on their approach—text-based, image-based, and multimodal—as well as their use of machine learning, deep learning, and pre-trained models. Research efforts have focused on binary vs. multi-label classification and bias mitigation.

Kiela et al. (2020) with the release of the Hateful Memes dataset, showed that unimodal architectures—text-only or image-only—consistently perform worse than multimodal ones. Their results provided the fundamental justification for combining textual and visual modalities in meme classification tasks.

Further studies have investigated more complex fusion techniques. Pramanick et al. (2021) presented MOMENTA, a multimodal framework that combines entity-enriched representations with cross-modal attention. The results have shown that attention-based fusion outperforms simple feature concatenation when multiple modalities provide complementary information. Our architecture was established directly on this design principle, using a cross-modal attention layer to allow text representations to attend over CLIP visual features rather than concatenating them.

Regarding the Nepali language, the MemeNePAL dataset was one of the first large collections of Nepali internet memes introduced by Thapa et al. (2025a) with sentiment and hate speech annotations. Multilingual BERT-based text-only models produce comparatively moderate Macro-F1 scores , while multimodal models consistently improve these scores, according to the study. This highlights the challenges of classifying meme sentiment in low-resource languages and the need for multimodal methods.

XLM-RoBERTa (Conneau et al., 2020), which has been pre-trained on 2.5TB of curated Common Crawl text in 100 languages using a SentencePiece BPE vocabulary, has stand out as the top general-purpose multilingual encoder for cross-lingual transfer tests such as XNLI, XQuAD, and MLQA. Its larger variant (with 24 layers, a hidden size of 1024, and 560M parameters) has significantly outperformed the base variant on low-resource languages like Nepali, where the increased model capacity offsets the limited pre-training signal available per language.

In vision-language modeling, CLIP (Radford et al., 2021) has utilized contrastive learning on a vast collection of image-text pairs to create aligned multimodal representations, whereas BLIP-2 (Li et al., 2023) has facilitated effective image caption generation by employing frozen encoders and lightweight transformers. We have used BLIP-2 to generate English-language visual descriptions of each meme, providing the text encoder with a natural-language rendering of the image content

that can be processed in the same token space as the extracted meme text. This design has allowed the model to reason jointly over image content and meme caption without requiring cross-lingual visual grounding.

Prior research further shows that domain-adaptive pre-training improves downstream performance in low-resource settings (Gururangan et al., 2020). Furthermore, methods for generating artificial code-switched text have been proposed to improve cross-lingual feature alignment and model performance on mixed-language tasks. We have incorporated these code-mixed samples into each training fold while deliberately withholding them from validation, ensuring that fold-level metrics remain comparable to the monolingual test distribution(Qin et al., 2020).[1]

3. Data

We have utilized the dataset introduced in prior work (Thapa et al., 2025a,b).The annotation schema follows CrisisHateMM (Bhandari et al., 2023). The dataset has been segmented into training, validation, and test sets containing 1,061, 133, and 133 samples, respectively. It primarily consists of Nepali memes written in Devanagari script, reflecting cultural and political discourse commonly observed in Nepali online communication. The dataset exhibits a class imbalance skewed toward the neutral class, as shown in Table 1.

Sets	Negative	Neutral	Positive	Total
Train	341	473	247	1,061
Validation	39	65	29	133
Test	–	–	–	133
Total	380	538	276	1,194

Table 1: Label Distribution for the Nepali Meme Dataset.

We have additionally incorporated the code-mixed split of the dataset as training augmentation data. This split contains 1,391 Nepali-English code-mixed meme images annotated for sentiment, as shown in Table 2. These samples have been used only during training and have been excluded from all validation splits.

Sets	Negative	Neutral	Positive	Total
Code-Mixed	263	757	371	1,391

Table 2: Label Distribution for the NeMeme Code-Mixed Augmentation Data.

[1] https://zenodo.org/records/15164380

4. Methodology

4.1. Data Preprocessing

4.1.1. Text Preprocessing

Text preprocessing has involved removing URLs, punctuation marks, and special characters to clean the input data. Then the cleaned text has been tokenized using XLM-RoBERTa SentencePiece vocabulary . Since the dataset comprises both monolingual Nepali and code-mixed Nepali–English text in Devanagari script, no transliteration has been applied; the multilingual tokenizer has handled both registers natively.

4.1.2. Image Preprocessing

We have converted all meme images to RGB format. In feature extraction, the images have been resized to 224×224 pixels and normalized to meet the requirements of the CLIP and BLIP-2 models. **Image-Text Similarity Scaling:** Following extracting and normalizing image embeddings,cosine similarity between image and corresponding text embeddings has been calculated. Negative similarity values has been clamped to zero, and the resulting scores were scaled by a factor of 2.5.

4.1.3. OCR Noise Filtering

EasyOCR extraction on code-mixed meme images frequently yields structurally corrupt output due to mixed Devanagari–English script boundaries and stylized meme fonts. Structural analysis of the 1,391 code-mixed OCR samples revealed that 2.9% (40 samples) contained problematic output as high-noise mixed text. Representative examples are shown in Table 3.

Category	Full OCR Output
Noisy	`%"[|=1-[1104; Meurlt "-|11[711`
Noisy	`Relar YE 5 coming Tike IIlrry * :llls :: 1-1!*170)/771!`
Noisy	`ce 9 "T_6 <o[+e dusltr l-ua tluzue u01=8 "=""371-`

Table 3: Examples of noisy OCR output from code-mixed meme images showing full EasyOCR output.

To investigate whether removing these samples improves model robustness, we experimented with three filtering strategies applied exclusively to the code-mixed augmentation data: (i) a Unicode range-based structural filter (Regex), which removes text with fewer than 50% valid Devanagari or Latin characters; (ii) pseudo-perplexity scoring (Salazar et al., 2020) using DistilBERT-Nepali (Sakonii, 2021), a model pre-trained on 13

million Nepali text sequences using masked language modeling, where texts exceeding an auto-calibrated perplexity threshold are removed; and (iii) LLM-based binary classification using Phi-3.5-mini (Abdin et al., 2024), which classifies each OCR text as CLEAN or NOISY. Results are presented in Table 4.

Filter	Scope	OOF F1
Baseline(no filtering)	–	**0.50**
Regex	Codemix only	0.47
DistilBERT-Nepali	Codemix only	0.44
Phi-3.5-mini	Codemix only	0.47

Table 4: Effect of OCR noise filtering strategies on system performance (XLM-R-large + CLIP + BLIP-2). OOF F1 = out-of-fold Macro-F1 on combined train and validation set.

None of the filtering strategies improved OOF F1 over the unfiltered baseline. We attribute this to the dual role of BLIP-2 captions: they provide a language-agnostic visual signal that independently compensates for failed or noisy OCR, while preserving all training samples and their visual representations through CLIP and BLIP-2. Filtering reduces available training data without proportional quality gains, as removed samples retain full visual signal regardless of OCR quality. This finding confirms that the multimodal architecture is inherently robust to OCR noise, rendering explicit filtering unnecessary for this task.

4.2. Data Augmentation

The training set has been augmented at the token level during training to improve generalization. With 25% probability, 15% of input tokens have been replaced with [MASK]; with 20% probability, 10% of words have been randomly dropped; and with 15% probability, adjacent word pairs have been randomly swapped. These stochastic perturbations have been applied only during training and only to the OCR-extracted meme text, not to the image captions. Additionally, 1,391 code-mixed Nepali–English meme samples have been incorporated as training augmentation, with monolingual samples upweighted by a factor of 2 via a WeightedRandomSampler to reduce reliance on the noisier code-mixed source.

4.3. Overview of Experimented Models

4.3.1. Unimodal Models

We have evaluated text-only baselines using three multilingual pre-trained encoders that have been fine-tuned end-to-end on the meme text column: XLM-RoBERTa-base (125M), MuRIL-base-cased (237M), and Twitter-XLM-RoBERTa-base-sentiment (125M). A linear classification head over the [CLS] token has been used. Then the models have been optimized with AdamW using a learning rate of $2e{-}5$, and have been trained for up to 10 epochs with early stopping. The architecture of the unimodal text model has been illustrated in Figure 2.

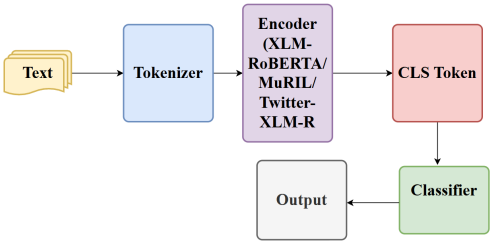

Figure 2: Unimodal Architecture for Text Data Processing and Classification.

4.3.2. Multimodal Models

Our multimodal system has fused three streams: (i) contextual text features from a fine-tuned XLM-RoBERTa-large applied to OCR-extracted meme text combined with BLIP-2 captions (Li et al., 2023); (ii) pooled visual features from a frozen CLIP ViT-L/14 encoder (Radford et al., 2021); and (iii) a BLIP-2 caption embedding along with a CLIP image–text cosine similarity scalar.

Fusion has been performed using a single-layer, 8-head cross-modal attention block with a hidden size of 1024 (128 per head), where the text [CLS] representation from XLM-RoBERTa-large attends over the CLIP image embedding as key and value. The attended text representation has then been added residually and normalized using LayerNorm, and subsequently concatenated with two 256-dimensional projected CLIP streams (image and caption) and a 32-dimensional similarity scalar. The resulting 1568-dimensional vector has passed through a two-layer MLP classifier with GELU activations and dropout ($p = 0.4$). The multimodal architecture has been illustrated in Figure 3.

The XLM-RoBERTa-large backbone (560M parameters) has been partially frozen: the embedding layer and the bottom 18 transformer layers have been frozen, leaving 76.6M trainable parameters. Two learning rates have been used, with a cosine schedule and 10% linear warmup. We have also experimented with NepaliBERT as an alternative text backbone, paired with the same frozen CLIP ViT-L/14 visual stream and fusion head.

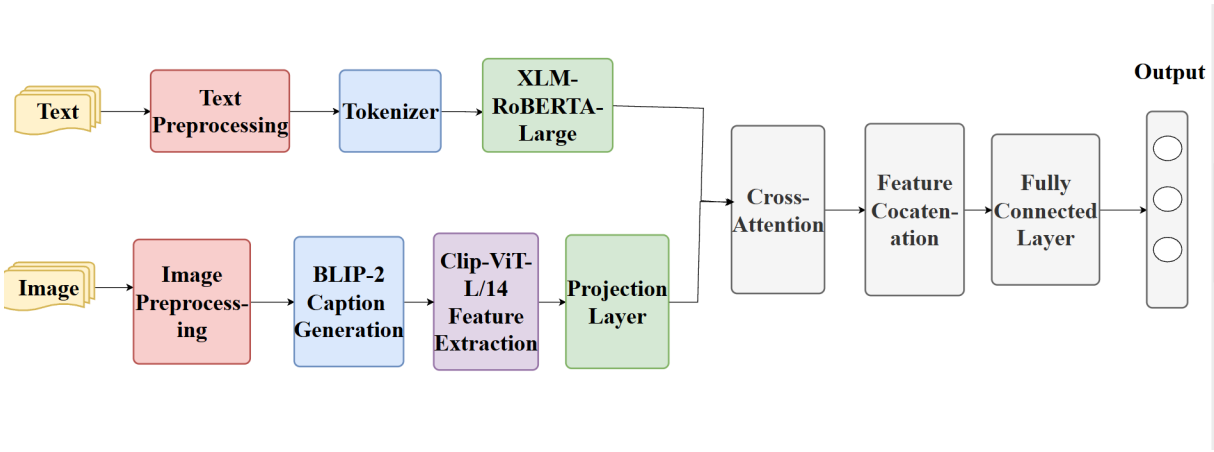

Figure 3: Multimodal Architecture using XLM-RoBERTa-Large and Clip-ViT-L/14.

5. Results and Analysis

This section presents the results of our Nepali meme sentiment classification task, comparing unimodal and multimodal approaches. We have evaluated model performance using macro-averaged precision, recall, and F1 score, with Macro-F1 serving as the primary evaluation metric.

5.1. Comparative Analysis

We have found that among unimodal text classifiers, muril-base-cased performed best, achieving a Macro-F1 of 0.42. It has outperformed xlm-roberta-base (0.40) and twitter-xlm-roberta-base-sentiment (0.39), highlighting the benefit of Nepali-inclusive pretraining.

For multimodal systems, the combination of xlm-roberta-large, CLIP ViT-L/14, and BLIP-2 has achieved the highest overall performance with a Macro-F1 of 0.50, improving +0.08 over the strongest text-only baseline. Specifically, incorporating BLIP-2 captions into the XLM-R-large + CLIP ViT-L/14 system has improved Macro-F1 from 0.42 to 0.50 (+0.08), as shown in Table 6, representing the single largest component-level gain in our ablation. This improvement is attributed to BLIP-2 providing an English-language visual description of each meme, compensating for noisy or empty OCR output in Devanagari script and supplying the text encoder with a language-agnostic rendering of the visual content that complements the raw image embeddings from CLIP.

Replacing cross-attention fusion with simple feature concatenation has reduced the mean OOF Macro-F1 from 0.4775 to 0.4438 (-0.0337), confirming that cross-modal attention is essential for effective modality interaction. Table 5 compares both fusion strategies across all folds.

Ablation studies demonstrate the contribution of each component: replacing CLIP ViT-L/14 with ViT-B/32 has reduced performance to 0.38, substituting XLM-R-large with NepaliBERT has yielded 0.46, and removing BLIP-2 captions has yielded 0.42. These results has confirmed that both aligned visual–textual embeddings and image captions play an important role in multimodal sentiment classification. Table 6 shows the performance of these models.

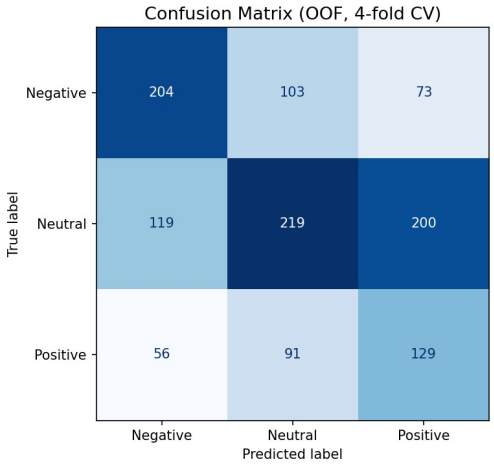

Figure 4: Confusion Matrix of the Multimodal XLM-R-large, CLIP and BLIP-2 model.

Metric	Cross-Attention	Concatenation
Fold 1 Val F1	0.48	0.41
Fold 2 Val F1	0.46	0.46
Fold 3 Val F1	0.51	0.48
Fold 4 Val F1	0.46	0.41
Mean OOF Macro-F1 $\pm$ Std	**0.4775 $\pm$ 0.0205**	**0.4438 $\pm$ 0.0307**

Table 5: Best Macro-F1 Scores for Each Fold and Mean OOF Macro-F1 Comparing Cross-Attention Fusion and Simple Concatenation in 4-Fold Stratified Cross-Validation on the Combined Training and Validation Dataset.

Classifier	Macro Average		
	P	R	F1
Unimodal (Text)			
MuRIL-base	0.44	0.45	0.42
XLM-R-base	0.43	0.41	0.40
Twitter-XLM-R	0.40	0.40	0.39
Multimodal			
XLM-R-base + CLIP ViT-B/32	0.42	0.44	0.38
NepaliBERT + CLIP ViT-L/14	0.46	0.46	0.46
XLM-R-large + CLIP ViT-L/14	0.42	0.44	0.42
XLM-R-large + CLIP + BLIP-2	**0.46**	**0.47**	**0.50**

Table 6: Performance of different systems on the test dataset.

5.2. Error Analysis

To further analyze model performance, we have presented the confusion matrix in Figure 4. It shows that our best model has correctly classified 204 Negative, 219 Neutral, and 129 Positive samples. However, substantial cross-class confusion has remained. For the Negative class (380 total), 103 instances have been misclassified as Neutral and 73 as Positive. Neutral (538 total) has been heavily confused with both extremes, with 119 predicted as Negative and 200 as Positive, which has led to its relatively low recall. Positive (276 total) has been the hardest class overall (F1 = 0.38), with 56 misclassified as Negative and 91 as Neutral, and a notably low precision due to frequent overprediction of Positive (402 total Positive predictions). Overall, the per-class F1 scores have followed the trend: Negative (0.53) > Neutral (0.46) > Positive (0.38). The confusion patterns have suggested strong semantic overlap between Neutral and the polar classes, while Positive has suffered most from class imbalance (23.1% of samples) and boundary ambiguity with Neutral.

Ablation results have further indicated that code-mixed augmentation has degraded performance when OCR quality has been poor. In our Nepali-Code-Mixed dataset, a subset of memes (2.9%) have produced empty or noisy text, which has weakened the text encoder's contribution and has increased misclassification—particularly between Neutral and Positive. The best-performing system has mitigated this issue by (i) incorporating BLIP-2 captions as a language-agnostic visual signal and (ii) upweighting monolingual samples using a WeightedRandomSampler, which has stabilized class learning without excessively biasing predictions toward the minority class. A systematic investigation of explicit OCR noise filtering strategies, detailed in Section 4.1.3, further confirmed that BLIP-2 already provides sufficient noise resilience, as no filtering approach improved over the unfiltered baseline.

5.3. Sensitivity to OCR Quality

The filtering ablation results presented in Section 4.1.3 indicate that the system is highly robust to OCR quality degradation. Structural analysis confirmed that only 2.9% of code-mixed OCR samples were genuinely problematic. The remaining semantic noise — OCR misreadings of colloquial Nepali meme text that appear structurally valid but are linguistically garbled — cannot currently be detected by automated methods without introducing false positives. The consistent failure of all three filtering strategies to improve OOF F1 further confirms that BLIP-2 visual captioning and WeightedRandomSampler together provide sufficient OCR noise resilience, making OCR quality a non-critical factor for this multimodal system.

6. Conclusion

In this research, we have evaluated unimodal and multimodal approaches for assessing the sentiments of Nepali memes. Multimodal systems consistently outperform text-only baselines, with our best system (XLM-RoBERTa-large + CLIP ViT-L/14 + BLIP-2, cross-modal attention fusion) achieving a Macro-F1 of 0.50 on the 4-fold OOF evaluation. A text-only MuRIL-large system with domain-specific warm-up on Nepali tweets achieves 0.4743, demonstrating that task-adaptive pre-training is a strong alternative to multimodal fusion when visual processing pipelines are unreliable. Future work will focus on Nepali-specific OCR, multilingual image captioning, and ensemble combination of the top-performing unimodal and multimodal systems.

7. Limitations

The dataset provided for our task is relatively small and imbalanced, particularly for the positive sentiment classes. Although code-mixed data augmentation has been incorporated, a subset of samples (2.9%) suffer from poor OCR quality, limiting the effectiveness of text features. Explicit OCR noise filtering strategies, including rule-based, transformer-based, and LLM-based approaches, did not improve over the unfiltered baseline, suggesting that more sophisticated noise detection methods tailored to colloquial low-resource meme text remain an open challenge. Furthermore the frozen visual backbone and limited text augmentation may not fully capture subtle multimodal cues, leading to some misclassifications.Additionally, our system has not employed end-to-end vision-language models (VLMs) such as LLaVA or InstructBLIP, which jointly reason over image and text in a unified framework and may better capture the nuanced visual-linguistic interactions present in meme con-

tent. Lastly, the performance of the model could be improved with more diverse and high-quality examples, especially for underrepresented memes.

8. Ethics Statement

In this study, we have developed our methodology following the highest ethical practices. By contributing to the identification of hate and sentiment understanding in low-resource memes, we hope to make the internet a safer and more inclusive place. We are committed to sharing our findings to prevent offensive and hateful content online while respecting linguistic and cultural diversity.

9. References

Marah Abdin et al. 2024. Phi-3 technical report. https://arxiv.org/abs/2404.14219.

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.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8342–8360.

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.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Thanh Tin Nguyen, Nhat Truong Pham, Ngoc Duy Nguyen, Hai Nguyen, Long H Nguyen, and Yong-Guk Kim. 2022. Hcilab at memotion 2.0 2022: Analysis of sentiment, emotion and intensity of emotion classes from meme images using single and multi modalities (short paper). In *DE-FACTIFY@ AAAI*.

Anil Singh Parihar, Surendrabikram Thapa, and Sushruti Mishra. 2021. Hate speech detection using natural language processing: Applications and challenges. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1302–1308. IEEE.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Momenta: A multimodal framework for detecting harmful memes and their targets. *arXiv preprint arXiv:2109.05184*.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. *arXiv preprint arXiv:2006.06402*.

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.

Sakonii. 2021. distilbert-base-nepali. https://huggingface.co/Sakonii/distilbert-base-nepali.

Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of ACL*.

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)*.

Limor Shifman. 2013. *Memes in digital culture*. MIT 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. 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.