linus@CHiPSAL 2026: Multimodal Hate Speech and Sentiment Detection in Low-Resource Memes using Late-Fusion Hybrid Architecture

Sunil Regmi[1], **Bipesh Subedi**[2], **Saugat Singh**[2], **Suman Shrestha**[2]
[1] Department of Artificial Intelligence, Kathmandu University
[2] Department of Computer Science and Engineering, Kathmandu University
sunilregmi233@gmail.com, bipeshrajsubedi@gmail.com, saugat.singh09@gmail.com, shrestha.suman@ku.edu.np

Abstract

The increased sharing of memes on social media creates serious challenges for automated moderation, especially in low-resource and code-mixed languages such as Nepali. In this paper, we present our system for the CHiPSAL 2026 Shared Task on Multimodal Hate and Sentiment Understanding in Low-Resource Memes. We propose a late-fusion hybrid architecture that combines OpenAI's Vision Transformer (CLIP ViT-B/32) with a domain-specific Nepali language model (NepBERTa) to capture both visual features and linguistic information. To address data scarcity, we introduce a cross-task label mapping and data augmentation strategy between the hate speech and sentiment datasets. By applying controlled hyperparameter settings and balanced loss optimization, our framework achieved a Macro F1 score of 0.8052 on Subtask A (Hate Speech Detection) and 0.6881 on Subtask B (Sentiment Analysis) in the official CodaBench evaluation, demonstrating the effectiveness of the proposed multimodal approach. The system was developed as part of a shared task under strict evaluation constraints where test labels were not publicly available.

Keywords: Multimodal, Hate Speech Detection, Sentiment Analysis, Code-Mixed Nepali, Late-Fusion

1. Introduction

Memes have become a distinctive and powerful form of digital communication, blending visual and textual information to convey a vast spectrum of emotions, opinions, and political critiques(Elahi et al., 2023a). As these digital constructs are deeply ingrained in internet culture, they serve as essential tools for sharing opinions and constructing collective identities across cultural boundaries(Konyspay et al., 2025). However, the rapid evolution of social media has also facilitated the proliferation of harmful content, such as hate speech, offensive trolling, and misinformation, which poses significant challenges for automated moderation systems(Mishra et al., 2023).

Understanding the core sentiment and intention behind memes is a complex task due to their frequent reliance on nuanced aggression, satire, and the complex relationship between words and visuals. Although significant studies have been carried out on memes in the English language, languages with fewer resources, such as Nepali, Hindi, Bengali and Tamil, have been investigated much less. These languages encounter specific challenges, such as a significant shortage of benchmark datasets, the absence of standardized optical character recognition (OCR) tools, and the common practice of using code-mixed or code-switched content (like "Nenglish," "Hinglish," "Banglish," or "Tanglish"), where native dialects are represented using English characters. To tackle these challenges effectively, research has progressively moved towards multimodal methods. Findings indicate that unimodal models, which examine text or images on their own, generally fall short as they overlook the contextual signals that arise from the interplay of both modalities.

Recent research shows that multimodal fusion, the collaborative combination of visual and textual elements greatly enhances the precision of sentiment analysis and trolling identification(Kannan and Rajalakshmi, 2022). Through the use of sophisticated deep learning models and Explainable AI (XAI) methods, researchers aim to create more resilient and understandable frameworks that elucidate the intricate social and emotional interactions of memes in linguistic contexts that lack resources(Elahi et al., 2023a).Recognizing these challenges in low-resource South Asian languages, the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) introduced a shared task focused on detecting hate speech and analyzing sentiment in Nepali-embedded images (Thapa et al., 2026; Sarveswaran et al., 2026). The task provides a structured benchmark for evaluating multimodal systems in code-mixed Nepali contexts, encouraging the development of robust approaches for harmful content detection in resource-constrained settings.

This paper outlines our team's methodological approach to hate speech detection and sentiment analysis on Nepali memes. Given the complexity

Proceedings of the Second workshop on Challenges in Processing South Asian Languages (CHiPSAL2026) @ LREC, pages 267–274

of standard computational social science analysis paradigms (Parihar et al., 2021; Thapa et al., 2025), we developed a structurally augmented Late-Fusion framework. By utilizing **NepBERTa** for high-fidelity textual extraction alongside **CLIP** for robust visual anchoring, we significantly bridge the localized context gap often encountered when analyzing text-embedded Nepali images. We also introduce a methodology for cross-task augmentation that maximizes the utility of sparse annotated data, proving that mapping implicit labels across overlapping domains provides significant stabilization during training.

2. Related Work

Multimodal hate speech and sentiment analysis in memes has evolved significantly over the past decade. Early approaches primarily focused on unimodal textual analysis using traditional machine learning algorithms such as Support Vector Machines (SVMs) and Naive Bayes classifiers, later transitioning to recurrent neural networks (RNNs) and standard Transformer architectures. While these methods achieved reasonable performance in text-only hate detection, they struggle in multimodal settings such as memes, where textual content may appear benign in isolation but becomes offensive when interpreted alongside visual context.

To address this limitation, subsequent research introduced multimodal fusion frameworks that combine textual and visual features. For instance, (Dubey et al., 2025) curated a dataset of 9,262 Hindi memes and employed OCR-based text extraction followed by multilingual embeddings (LASER, LaBSE) and CNN-based image encoders (e.g., VGG16, ResNet18), achieving 80% accuracy through late fusion. Similarly, (Karim et al., 2022) extended the Bengali Hate Speech Dataset with 4,500 labeled memes and demonstrated that integrating CNN backbones with transformer-based language models (RoBERTa, XLM-R, BanglaBERT) improves performance, achieving an F1-score of 0.79 and MCC of 0.808. (Rajput et al., 2022) further explored code-switched political memes using a CNN–LSTM multimodal framework, reporting an F1-score of 0.792 and highlighting the challenges of multilingual and politically nuanced content.

More recent advancements reflect a paradigm shift toward joint multimodal representation learning through large-scale vision–language pretrained models. Architectures such as VisualBERT, LXMERT, and CLIP align textual and visual embeddings in a shared latent space, enabling cross-modal reasoning without extensive task-specific feature engineering. A CLIP-based architecture with prompt engineering achieved 87.42% accuracy and 90.13% F1-score in a zero-shot hate speech detection setting (Arya et al., 2024). In the context of multimodal hate speech event detection, (El-Sayed and Nasr, 2024) combined CLIP with transformer-based text models such as HateBERT, RoBERTa, and XLM-R on the CrisisHateMM dataset, achieving an F1-score of 0.7703.

Parallel efforts in multimodal sentiment analysis have shown similar trends. (Elahi et al., 2023b) introduced the MemoSEN dataset of 4,372 Bengali memes categorized as positive, negative, or neutral, using ResNet50 and BanglishBERT to achieve a weighted F1-score of 0.71. Similarly, (Guleria et al., 2024) analyzed English and Hinglish memes using a combination of RoBERTa and CLIP, reporting an accuracy of 0.82.

Despite these advances, multimodal hate speech detection in South Asian languages remains relatively underexplored, particularly for low-resource languages such as Nepali. Existing studies emphasize the importance of localized language models for handling code-switched, morphologically rich, or transliterated text, as generalized multilingual models may inadequately capture linguistic nuances. Motivated by these gaps, we adopt a hybrid paradigm that integrates the generalization strength of vision–language pretraining with the linguistic specificity of a regional transformer. Specifically, we fuse the visual backbone of CLIP with NepBERTa, a RoBERTa-based model tailored for Nepali language understanding, enabling robust multimodal representation learning in low-resource meme datasets.

3. Dataset & Task

The primary data used is the **CHiPSAL 2026 Dataset**, which compiles low-resource, code-mixed Nepali memes. The dataset draws from foundational multimodal datasets and task architectures in the domain (Thapa et al., 2025a,b; Bhandari et al., 2023). Due to the linguistic diversity of South Asia, the dataset inherently contains heavy code-switching - primarily a mixture of Devanagari script, phonetically transliterated English written in Devanagari, and English terms written natively. This structural complexity introduces substantial challenges for standard parsers. The shared task is divided into:

- **Subtask A (Hate Speech Detection):** A binary classification task to identify whether a meme contains "Hate" (label 1) or "No-Hate" (label 0). Hate speech in this context constitutes explicit slurs, targeted attacks, violence incitement, or profoundly offensive rhetoric targeting marginalized groups based on race, gender, religion, or caste. Determining this

often requires inferring the intent behind the overlaying text when combined with an inciting conceptual image.

- **Subtask B (Sentiment Analysis):** A three-way classification task categorizing memes into "Negative" (label 0), "Neutral" (label 1), or "Positive" (label 2). This task maps the generalized emotional polarity of the meme. Note that positive sentiment imagery overlaid with highly sarcastic text frequently maps to a "Negative" ground truth, demanding deep multi-task analytical capabilities.

3.1. Cross-Task Data Creation and Augmentation

To address data scarcity and improve training robustness, we explore a cross-task data augmentation strategy that leverages the overlap between sentiment polarity and hate speech signals in meme data. However, this relationship is not always straightforward. In meme contexts, particularly in low-resource and code-mixed settings, sentiment labels can be ambiguous. For example, samples labeled as "Positive" may sometimes express sarcasm or irony, which can align with underlying hateful intent depending on context. Therefore, the connection between sentiment and hate labels should be treated as approximate and context-dependent rather than strictly equivalent. We highlight this as a potential direction that requires more careful investigation in future work.

In this work, we adopt a simple heuristic-based cross-task label mapping to construct augmented datasets. Specifically, we approximate the alignment by treating negative sentiment as indicative of hate and positive sentiment as indicative of non-hate. While this simplification does not capture all nuances, it provides a practical way to expand the training data. The detailed statistics for both subtasks, including original, augmented, and validation splits, are reported in Tables 1 and 2.

- **Augmenting Subtask A:** We extend the Subtask A dataset using samples from Subtask B. Since Subtask A is a binary classification task, we first remove all “Neutral” (1) instances from Subtask B. The remaining samples are mapped as follows: Subtask B “Negative” (0) $\rightarrow$ Subtask A “Hate” (1), and Subtask B “Positive” (2) $\rightarrow$ Subtask A “No-Hate” (0).
- **Augmenting Subtask B:** Similarly, we augment Subtask B using samples from Subtask A by projecting binary labels onto the sentiment space. Specifically, Subtask A “No-Hate” (0) is mapped to Subtask B “Positive” (2), while Subtask A “Hate” (1) is mapped to Subtask B “Negative” (0).

Split	Total	No Hate	Hate
train	1068	348	720
train(augmented)	1656	595	1061
validation	133	35	98

Table 1: Subtask A (Hate Speech Detection) Dataset Statistics

Split	Total	-ve	Neutral	+ve
train	1061	341	473	247
train(aug.)	2129	1061	473	595
validation	133	39	65	29

Table 2: Subtask B (Sentiment Analysis) Dataset Statistics

4. Methodology

4.1. Data Preprocessing

Before inputting the data streams into the multimodal architecture, rigorous pre-processing pipelines were established for both textual and visual modalities.

Images were first normalized according to CLIP prerequisites, which enforce specific channel-wise mean and standard deviation metrics to align the visual geometry to the original pre-training subspace. To establish extreme robustness against standard formatting variations seen in low-resolution internet memes, dynamic PyTorch visual augmentations were applied algorithmically bounding the training datasets. We used a combination of `RandomResizedCrop` (to scale against shifting aspect ratios), `RandomHorizontalFlip`, `RandomRotation` up to 15° (to counter angled text constraints), and aggressive `ColorJitter` (dynamically adjusting brightness, contrast, and saturation variations common in recaptured media).

Text elements native to the imagery were tokenized using the `Rajan/nepbertaTorch`[1] Subword Tokenizer. Standard generic preprocessing approaches often employ basic whitespace delimitation which catastrophically breaks morphological Nepali structures. Our tokenizer preserves localized contextual information, encoding sub-words efficiently. Truncation was enforced at a static sequence length limit of 128 tokens, zero-padding the subsequent sequence fields. This uniform bounding was essential for scaling parallel processing during the highly constrained hyperparameter tuning phase.

4.2. Model Architecture

To ensure accurate localization against native Nepali dialects while maintaining robust visual con-

[1] https://huggingface.co/Rajan/nepbertaTorch

text, our core modeling strategy leverages a **Hybrid Late-Fusion** architecture matching a specialized text extraction module with a broad-domain visual foundation model.

4.2.1. Visual Feature Extraction

We utilize OpenAI's CLIP (ViT-B/32) (Radford et al., 2021) to extract high-quality visual representations. CLIP's Vision Transformer parses the input image into a sequence of non-overlapping patches, which are linearly embedded and prepended with a learnable classification (`[CLS]`) token. The final hidden state corresponding to this `[CLS]` token serves as the global visual representation, denoted as $v \in \mathbb{R}^{d_v}$, where $d_v = 512$. By leveraging CLIP, the model inherently possesses strong zero-shot capabilities and robust structural understanding of diverse, internet-native imagery, which is crucial for meme analysis.

4.2.2. Textual Feature Extraction

For the textual modality, we employ **NepBERTa**, a transformer model specifically pre-trained on a vast corpus of Nepali text. Traditional multilingual models like mBERT or Multilingual-CLIP often use byte-pair encodings that poorly align with Devanagari script, resulting in fragmented subwords and diluted semantic representation (Thapa et al., 2025). NepBERTa, conversely, captures the morphological intricacies of Nepali. The tokenized input sequence is passed through the transformer layers, and the embedding corresponding to the sequence-level `[CLS]` token is extracted as the text representation $t \in \mathbb{R}^{d_t}$, where $d_t = 768$.

4.2.3. Late-Fusion and Classification

The core of our approach lies in the late-fusion mechanism. Premature fusion can lead to one modality dominating the gradient updates, especially when textual cues (like explicitly offensive words) provide stronger immediate signals than complex visual context. To prevent this, the extracted features are independently L2-normalized to ensure stable gradient scales:

$$\hat{v} = \frac{v}{\|v\|_2}, \quad \hat{t} = \frac{t}{\|t\|_2} \tag{1}$$

These normalized embeddings are then concatenated into a dense 1280-dimensional joint representation z:

$$z = \hat{v} \oplus \hat{t} \in \mathbb{R}^{1280} \tag{2}$$

To synthesize and stabilize the multimodal gradients before classification, a `LayerNorm` operation is applied to z. The fused state then passes through a dense multi-layer perceptron (MLP) structured as follows:

1. Linear transformation reducing dimensionality from 1280 to 512, followed by a GELU activation and Dropout (0.3).
2. Linear transformation from 512 to 128, followed by a GELU activation and Dropout (0.2).
3. Final Linear projection from 128 to the target num_labels, producing the output logits $\hat{y}$.

4.3. Training Setup

Training optimizations strictly leveraged adaptive `AdamW` implementations. Realizing the danger of catastrophic forgetting in pre-trained instances, we deployed significantly lower learning rates for the backbones (2×10^{-6} to 3×10^{-6}) while implementing a steeper gradient curve for the classifier head (1×10^{-4} to 2×10^{-4}), managed via a `OneCycleLR` scheduler.

Furthermore, we instituted aggressive explicit class weights (e.g., `[1.0, 4.0, 1.0]`) to handle "Neutral" label sparsity in Subtask B. Additionally, `label_smoothing=0.1` was integrated to throttle overconfidence biases favoring major label instances. Hardware parameters utilized active Gradient Accumulation to guarantee high generalized batch-resolutions.

4.4. Hyperparameter Optimization and Setup

To discover the optimal learning configuration that balances the pre-trained weights without causing catastrophic forgetting, we conducted structured grid searches. Our optimization tracked parameter subsets varying the backbone learning rates from 1×10^{-6} to 5×10^{-6} and head learning rates from 5×10^{-5} up to 2×10^{-4}. We implemented a mini-batch processing logic mapping target batch sizes to dynamic `gradient_accumulation_steps`, ensuring consistent optimization geometries extending beyond standard VRAM limitations. The final models were evaluated after 50 epochs utilizing an `AdamW` optimizer scaled harmoniously via the `OneCycleLR` scheduler.

5. Results & Discussion

Task	Acc.	F1 Mac.	Prec.	Recall
Sub. A	0.8209	**0.8052**	0.7974	0.8202
Sub. B	0.6842	**0.6881**	0.6979	0.7013

Table 3: Performance on Official Test Benchmarks

The models were evaluated using the official CodaBench evaluation system, where performance

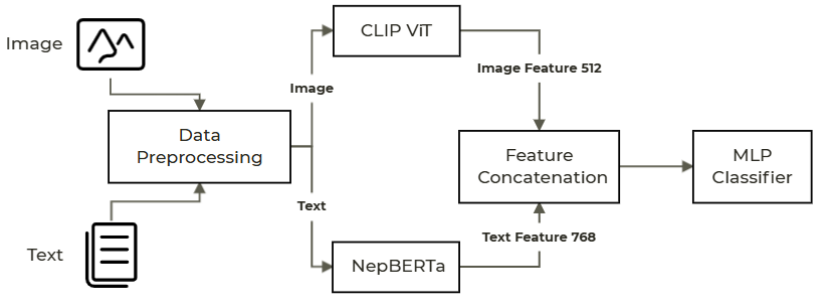

Figure 1: Overall architecture of the proposed multimodal framework combining CLIP-based visual features with NepBERTa textual embeddings using late fusion and regularized classification layers.

was measured in terms of Accuracy, Macro-F1, Precision, and Recall on the respective test sets. Detailed results are reported in Table 3.

Integrating cross-task augmented data into Subtask A allowed us to explicitly amplify the boundaries between Hate and No-Hate classes. By completely removing Neutral instances from the augmentation process, we avoided introducing ambiguous signals, which contributed to Subtask A achieving a peak Test F1 of 0.8052. This confirms the effectiveness of a constrained cross-task augmentation strategy that leverages overlapping domains while controlling for label noise.

In contrast, Subtask B remains more challenging due to its three-class structure. While Subtask A benefits from explicit binary boundaries, Subtask B requires distinguishing among Positive, Negative, and Neutral classes. Even though Neutral instances were not added via augmentation, the presence of this intermediate class inherently introduces ambiguity, as subtle expressions and code-mixed nuances often overlap with both Positive and Negative labels. This likely contributes to the relatively lower Macro F1 score of 0.6881 compared to Subtask A.

To mitigate overfitting and improve generalization, techniques such as `label_smoothing` and dropout were applied. These methods were particularly helpful for Subtask B, where noisy textual signals from memes can otherwise cause the model to memorize the training data. Interestingly, the Subtask B model achieved a Macro F1 score on the official test set that was higher than its internal validation score (0.6881 vs. 0.6607), demonstrating improved generalization to unseen examples.

During hyperparameter tuning, we observed that setting the learning rate too high for the pre-trained backbones led to drastic performance drops. When backbone learning rates exceeded 2×10^{-6}, F1 scores fell to around 0.55, likely because the pre-trained representations in NepBERTa and CLIP were disrupted. The best performance was achieved with a backbone learning rate of 3×10^{-6} and a higher learning rate of 2×10^{-4} for the classification head. Keeping backbone learning rates low allowed the classification head to effectively learn feature fusion without destroying pre-trained knowledge.

While the model achieves competitive performance, there remains room for improvement, particularly in handling complex multimodal interactions where the relationship between text and visual content is complex or context-dependent. These challenges are inherent to the task and highlight potential directions for future work.

5.1. Ablation Study

To better understand the contribution of each component, we perform ablation experiments on the validation split as shown in Table 4. Due to shared task constraints, ground-truth labels for the official test set are not publicly available. Therefore, all comparative experiments are conducted using validation data, which was also used for model selection prior to final submission.

Model	Subtask A (F1)	Subtask B (F1)
CLIP-only	0.54	0.59
NepBERTa-only	0.63	0.60
Fusion (Ours)	0.78	0.66

Table 4: Ablation study on validation set

The results show that the fusion model outperforms both unimodal variants across both subtasks. CLIP-only achieves the lowest scores, indicating that visual features alone are insufficient for accurate meme understanding. NepBERTa-only performs better, suggesting textual information plays a stronger role. However, the fusion model achieves the highest performance, improving Subtask A from 0.63 to 0.78 and Subtask B from 0.60 to 0.66.

This confirms that combining visual and textual features helps capture additional context unavailable to single-modality models. The observation aligns with prior multimodal learning research,

where integrated vision–language representations improve tasks requiring joint reasoning over text and images (Dubey et al., 2025; Karim et al., 2022). Recent aligned vision-language models further support the efficacy of multimodal fusion, demonstrating strong generalization across diverse tasks.

Overall, the results highlight the importance of multimodal fusion and controlled cross-task augmentation for complex meme datasets, especially in low-resource and code-mixed settings. The final fusion model was selected based on validation performance and submitted to the official evaluation server, where it achieved the reported test scores.

6. Conclusion

This work demonstrates that combining a domain-specific language model (NepBERTa) with a strong visual encoder (CLIP) provides an effective baseline for multimodal understanding in low-resource, code-mixed meme data. The proposed late-fusion approach achieved competitive performance in the CHiPSAL 2026 shared task. Our findings suggest that cross-task augmentation can provide additional training signals in low-data settings, although its effectiveness depends on the quality of label mapping. The results also highlight the importance of textual features, while visual features provide complementary context.

Future work can explore improved OCR systems, better handling of sarcasm, and more advanced multimodal fusion techniques.

Data and Model: This work utilizes the CHiPSAL 2026 shared task dataset, publicly available at https://github.com/therealthapa/chipsal26-memes . The pretrained model used in this study are available in hugging face. The implementation code and the models will be made publicly available in a repository (https://github.com/sunilRegmi-ai/chipsal2026) following publication .

7. Limitations

A key limitation of our work is that ablation studies could not be performed on the official test set, as ground-truth labels were not publicly available during the competition. Although we report unimodal and fusion comparisons on the validation set, the exact contribution of each component on the test set remains uncertain. The proposed cross-task label mapping is heuristic and assumes a direct relationship between sentiment and hate speech (e.g., positive sentiment mapped to non-hate and negative sentiment mapped to hate). While this assumption is intuitive, it may not always hold in practice, particularly in cases such as sarcasm, implicit hate, or context-dependent expressions, which may introduce noisy or ambiguous training signals.While we provide ablation results comparing unimodal and fusion models, we did not isolate the effect of cross-task augmentation separately. Therefore, its exact contribution to performance improvement cannot be fully quantified. We also did not experimentally evaluate alternative fusion strategies such as early fusion or cross-attention. Our choice of late fusion is based on simplicity and stability rather than direct empirical comparison with other fusion methods. In addition, the system depends on OCR-extracted text, making it sensitive to OCR errors and degraded image quality. When text extraction fails or produces incorrect tokens, model performance can degrade significantly.

Finally, the model struggles with sarcasm and culturally nuanced expressions, which are common in meme data. Addressing these challenges would require more advanced multimodal reasoning and improved language understanding in low-resource settings.

8. Ethics Statement

This work uses the CHiPSAL 2026 shared task dataset, which is publicly available for research. Meme-based data may contain sensitive or biased content, and models trained on such data can reflect or amplify these biases. The model also relies on pretrained representations from the Hugging Face ecosystem, which may introduce additional biases.

In terms of limitations, the model may not generalize well to unseen or culturally diverse memes and does not explicitly address bias mitigation, robustness, or interpretability.

9. Bibliographical References

Greeshma Arya, Mohammad Kamrul Hasan, Ashish Bagwari, Nurhizam Safie, Shayla Islam, Fatima Rayan Awad Ahmed, Aaishani De, Muhammad Attique Khan, and Taher M. Ghazal. 2024. Multimodal hate speech detection in memes using contrastive language-image pre-training. *IEEE Access*, 12:22359–22375.

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*.

Ahmed El-Sayed and Omar Nasr. 2024. Aast-nlp at multimodal hate speech event detection 2024

: A multimodal approach for classification of text-embedded images based on clip and bert-based models. In *CASE*.

Kazi Toufique Elahi, Tasnuva Binte Rahman, Shakil Shahriar, Samir Sarker, Sajib Kumar Saha Joy, and Faisal Muhammad Shah. 2023a. Explainable multimodal sentiment analysis on bengali memes.

Kazi Toufique Elahi, Tasnuva Binte Rahman, Shakil Shahriar, Samir Sarker, Sajib Kumar Saha Joy, and Faisal Muhammad Shah. 2023b. Explainable multimodal sentiment analysis on bengali memes. In *Proceedings of the International Conference on Computer and Information Technology (ICCIT)*.

Aishvi Guleria, Kamya Varshney, Garima Pahwa, Shreya Singhal, and Nonita Sharma. 2024. Multimodal sentiment analysis of english and hinglish memes. *Multimedia Tools and Applications*.

Ramesh Kannan and Ratnavel Rajalakshmi. 2022. Multimodal code-mixed Tamil troll meme classification using feature fusion. In *Proceedings of the First Workshop on Multimodal Machine Learning in Low-resource Languages*, pages 1–8, IIIT Delhi, New Delhi, India. Association for Computational Linguistics.

Md. Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Md. Shajalal, and Bharathi Raja Chakravarthi. 2022. Multimodal hate speech detection from bengali memes and texts. In *Proceedings of the International Conference on Speech and Language Technologies for Low-Resource Languages*.

Aidos Konyspay, Pakizar Shamoi, Malika Ziyada, and Zhusup Smambayev. 2025. Meme similarity and emotion detection using multimodal analysis. In *2025 International Conference on Activity and Behavior Computing (ABC)*, page 1–10. IEEE.

Shreyash Mishra, S Suryavardan, Parth Patwa, Megha Chakraborty, Anku Rani, Aishwarya Reganti, Aman Chadha, Amitava Das, Amit Sheth, Manoj Chinnakotla, Asif Ekbal, and Srijan Kumar. 2023. Memotion 3: Dataset on sentiment and emotion analysis of codemixed hindi-english memes.

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.

Kshitij Rajput, Raghav Kapoor, Kaushal Rai, and Preeti Kaur. 2022. Hate me not: Detecting hate inducing memes in code switched languages. *arXiv preprint arXiv:2204.11356*.

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, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. 2025. Large language models (llm) in computational social science: prospects, current state, and challenges. *Social Network Analysis and Mining*, 15(1):1–30.

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)*.

10. Language Resource 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.

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.