NeuralNoodles@CHiPSAL 2026: Late-Fusion Multimodal Stacking for Nepali Meme Sentiment Classification

Sidratul Muntaha[1], **Sabila Anzum**[1], **Arpita Mallik**[1], **Hasan Murad**[1]
[1]Chittagong University of Engineering & Technology (CUET), Chittagong, Bangladesh
{u2104119, u2104062, u2004023}@student.cuet.ac.bd, hasanmurad@cuet.ac.bd

Abstract

Memes have emerged to be an essential medium of online expression, where the sentiment is determined by the interaction of text and image. Sentiment analysis of memes is particularly challenging when the language is low-resource, such as Nepali, due to the lack of resources and the complex relationships between text and image modalities. In this paper, we report our submission to Subtask B of CHiPSAL 2026, where the task was sentiment analysis of Nepali text-embedded memes for three sentiment classes: Negative, Neutral, and Positive. Through this submission, we present a late fusion multimodal framework that encompasses lexical, semantic, and visual models through a cross-validated stacking approach. Our submission to the shared task competition received a Macro F1 of 0.5045 on the official test set, achieving 6th place in the leaderboard. This demonstrates the strength of well-structured late fusion approaches to multimodal sentiment analysis of text-embedded memes.

Keywords: Multimodal Sentiment Analysis, Nepali Memes, Low-Resource NLP, Late Fusion, Meme Classification, Deep Learning

1. Introduction

Memes have become a popular way for people to communicate online. They allow individuals to express complex emotions, opinions, and viewpoints in an engaging and brief way Guleria et al. (2025). Automatic understanding of memes can give us valuable insights into public sentiment, cultural attitudes, and societal reactions to events Bucur et al. (2022). Unlike traditional text-only sentiment analysis, meme sentiment comes from the interplay between visual and textual elements, not from either part alone. This combination makes classifying meme sentiment a difficult research challenge.

While there has been significant progress in multimodal learning for English and other widely used languages, low-resource languages like Nepali are still largely unexplored. Identifying the sentiment in memes is important for understanding public opinion and for reducing the spread of negativity, misinformation, and harmful narratives Hossain et al. (2022).

Recent studies highlight the benefits of multimodal approaches. For example, Hossain et al. (2022) introduced MemoSen, a Bengali multimodal meme dataset and demonstrated that combining image and text features significantly improves classification performance over unimodal systems. Similarly, Jha et al. (2025) found that transformer-CNN hybrid architectures have also shown great performances in meme sentiment analysis and other related tasks. However, systematic approaches for multimodal sentiment analysis of Nepali memes are scarce, especially for structured fusion approaches that focus on robustness and generalization.

To fill this research gap, we have presented NeuralNoodlesCHiPSAL 2026, a late-fusion multimodal framework for Nepali meme sentiment analysis. The main contributions of this work are as follows:

1. To build a hybrid multimodal framework for Nepali meme sentiment analysis.
2. To design a structured probability-level late fusion strategy using cross-validated stacking.
3. To provide a reproducible and computationally efficient baseline for low-resource multimodal settings.

Although developing advanced cross-attention transformer models is beyond the scope of our proposed work, our proposed stacking-based multimodal model is a strong and reproducible baseline for future multimodal sentiment classification of Nepali memes. The implementation of our proposed method is publicly available at: `https://github.com/sidratul-m00ntaha/NeuralNoodles-CHiPSAL-2026`.

2. Related Work

Recent research has closely examined multimodal sentiment analysis of memes in different languages and tasks. Hossain et al. (2022) introduced a Bengali multimodal meme dataset with 4,368 annotated samples. They have showed that combining text and visuals improves performance by about 1.2% compared to unimodal baselines. Many transformer-based multimodal systems did very well on the Memotion shared tasks. Bucur et al. (2022) has used BERT, Sentence Transformers,

Proceedings of the Second workshop on Challenges in Processing South Asian Languages (CHiPSAL2026) @ LREC, pages 250–257

EfficientNet, and CLIP within a multi-task transformer framework, securing top ranks in sentiment and humor tasks. Similarly, Messina et al. (2021) proposed a Double Visual Textual Transformer (DVTT) architecture. This architecture uses mutually conditioned transformers for images and text, merging probability outputs by averaging. Sharma et al. (2020) has applied transfer learning with CNN backbones and stacked BiLSTM-GRU attention mechanisms for multimodal fusion. Swamy et al. (2020) has explored both unimodal and bimodal transformer-based architectures, pointing out challenges in effectively using visual cues.

Several studies have proposed fusion and hybrid architectures beyond shared tasks. Jha et al. (2025) evaluated DEIT+SBERT, DistillBERT+ResNet, RoBERTa+ResNet, and BiLSTM+ResNet, reporting 78.83% accuracy with DistillBERT+ResNet. Guleria et al. (2025) combined RoBERTa and CLIP for Hinglish meme sentiment classification, achieving 0.82 accuracy and significantly outperforming traditional machine learning methods. Vankov et al. (2025) created a dataset of 5,592 samples using distant supervision. They demonstrated that multimodal approaches and large language models do better than unimodal systems. Sharma et al. (2024) proposed ALFRED, an emotion-aware cross-modal gating framework that improved the F1-score by 4.94% over the baselines. In domain-specific detection tasks, Roy et al. (2024) used Vision Transformers and pretrained text transformers to detect Bangla cyberbullying memes, achieving a 0.76 F1-score. , Meanwhile, Shanmugavadivel et al. (2025) applied hybrid CNN-transformer architectures for detecting misogyny in Tamil memes. Hasan et al. (2022) explored deep learning techniques for detecting multimodal troll memes in Tamil. They focused on visual and textual feature extraction using VGG16, ResNet50, VGG19, CNN, and CNN+LSTM architectures. Their multimodal framework combined VGG16 image features with CNN and LSTM-based textual representations. The textual CNN+LSTM model achieved the highest weighted F1-score (0.52) and recall (0.57). The CNN-Text+VGG16 setup has reached the highest multimodal F1-score of 0.49, allowing the team to secure 4th place in the shared task. Velioglu and Rose (2020) used VisualBERT with ensemble learning to gain a 0.811 AUROC. Additionally, Elahi et al. (2023) approaches using ResNet50 and BanglishBERT showed promising results (0.71 weighted F1) for Bengali memes, highlighting the need for interpretability. Together, these studies show that multimodal transformer-based fusion, transfer learning, and ensemble strategies greatly improve meme classification performance. However, most research focuses on high-resource or better-resourced languages. Structured stacking-based late fusion strategies for Nepali meme sentiment analysis are still underexplored.

Motivated by existing multimodal sentiment analysis methods, we have proposed a hybrid late-fusion approach to handle sentiment classification in Nepali memes.

3. Dataset and Task

3.1. Shared Task Description

In this study, we have used the dataset provided in the Shared Task on Multimodal Hate and Sentiment Understanding in Low-Resource Memes, held in CHiPSAL 2026 Thapa et al. (2026), where the focus is on the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2026) Sarveswaran et al. (2026). The shared task is about multimodal content moderation in low-resource languages, specifically Nepali memes.

The shared task is divided into two subtasks:

1. Hate Speech Detection in Nepali Only Memes

2. Sentiment Analysis in Nepali Only Memes

In this study, we have taken part in Subtask B, titled *Sentiment Analysis in Nepali Only Memes*, where the objective is to classify the memes according to the following sentiments: Negative, Neutral, and Positive.

3.2. Dataset Source and Construction

The dataset for this shared task was developed by extending the previous research works. To begin with, a multimodal framework for analyzing low-resource code-mixed memes was presented by Thapa et al. (2025a). This was followed by cross-platform retrieval and distillation techniques for improving the quality of the dataset, as presented by Thapa et al. (2025b). The annotation guidelines for detecting multimodal hate content were obtained from the CrisisHateMM framework, as presented by Bhandari et al. (2023)

3.3. Dataset Structure

The dataset is comprised of Nepali text-embedded meme image data. Table 1 presents the class-wise distribution of each set. Out of 1061 memes, 247 and 341 memes are from positive and negative classes respectively, while 473 are from the neutral class.

Class	Train	Test
Positive	247	–
Negative	341	–
Neutral	473	–
Total	1061	133

Table 1: Class-wise Distribution of the Dataset

4. Methodology

We have proposed a late-fusion multimodal framework for the sentiment classification of Nepali text-embedded memes (Subtask B). Due to the small dataset (∼1k memes) and the complexity of multimodal reasoning, we have proposed a modular design that combines lexical, semantic, and visual representations using a stacking ensemble method.

Our work differs from other state-of-the-art transformer-based multimodal models such as the Double Visual Textual Transformer (DVTT) model proposed by Messina et al. (2021), the Visual BERT model used in Velioglu and Rose (2020) and the gated cross-attention model proposed by Sharma et al. (2024) in the sense that our model focuses on stability and generalization on a low-resource dataset. The overall architecture of the proposed framework is presented in Figure 1.

4.1. Problem Formulation

For a given meme

$$M = (I, T),$$

where I denotes the image and T denotes the OCR-extracted textual content, the goal is to predict a sentiment label:

$$y \in \{0, 1, 2\},$$

which represents:

- 0: Negative
- 1: Neutral
- 2: Positive

The training set has a moderate level of class imbalance, with the neutral class being the majority class. To evaluate the model on all classes equally, the model's performance has been evaluated using **Macro F1 score**, which gives equal importance to each class compared to accuracy score.

4.2. Data Preprocessing

Before feature extraction, some basic preprocessing steps were applied to the text as well as the image data.

For text:

- Removal of URLs and excessive whitespace
- Normalization of punctuation
- Lowercasing where appropriate
- Retention of Nepali Unicode characters

For images:

- All images were resized to 224 × 224 resolution
- Pixel normalization was applied using ImageNet statistics

These preprocessing steps ensure consistency across modalities while preserving sentiment-relevant features.

4.3. Textual Representation

For textual data representation, we have used two complementary representations:

4.3.1. Lexical features

We have used a hybrid representation, referred to as TF-IDF, that includes a bag of word features comprising word-level (1-2 gram) and character-level (3-5 gram) n-grams. The inclusion of character features is to account for variant spellings and informal language often utilized in memes. For classification, we have utilized a Logistic Regression model with balancing classes

4.3.2. Semantic features

We have used a multilingual Sentence Transformer model, referred to as paraphrase-multilingual-MiniLM-L12-v2, that produced embeddings of dimensionality 384 for each piece of text. The model was frozen to prevent potential overfitting. We chose it due to its good performance on multilingual data. Besides, its lightweight architecture makes it well-suited for low-resource settings compared to larger models such as XLM-RoBERTa or MuRIL. We have utilised a Logistic Regression classifier for this model. This dual modelling captured both surface-level sentiment cues and contextual semantics.

For the TF-IDF case, the word-level vectorizer had an n-gram range of (1,2) with a maximum vocabulary of 3000 features, while the character-level vectorizer had an n-gram range of (3,5) with a maximum of 2000 features. The final sparse representation was obtained by combining the word and character TF-IDF matrices, resulting in a feature space of up to 5000 dimensions. Logistic Regression was used with class balancing and a maximum of 1000 iterations.

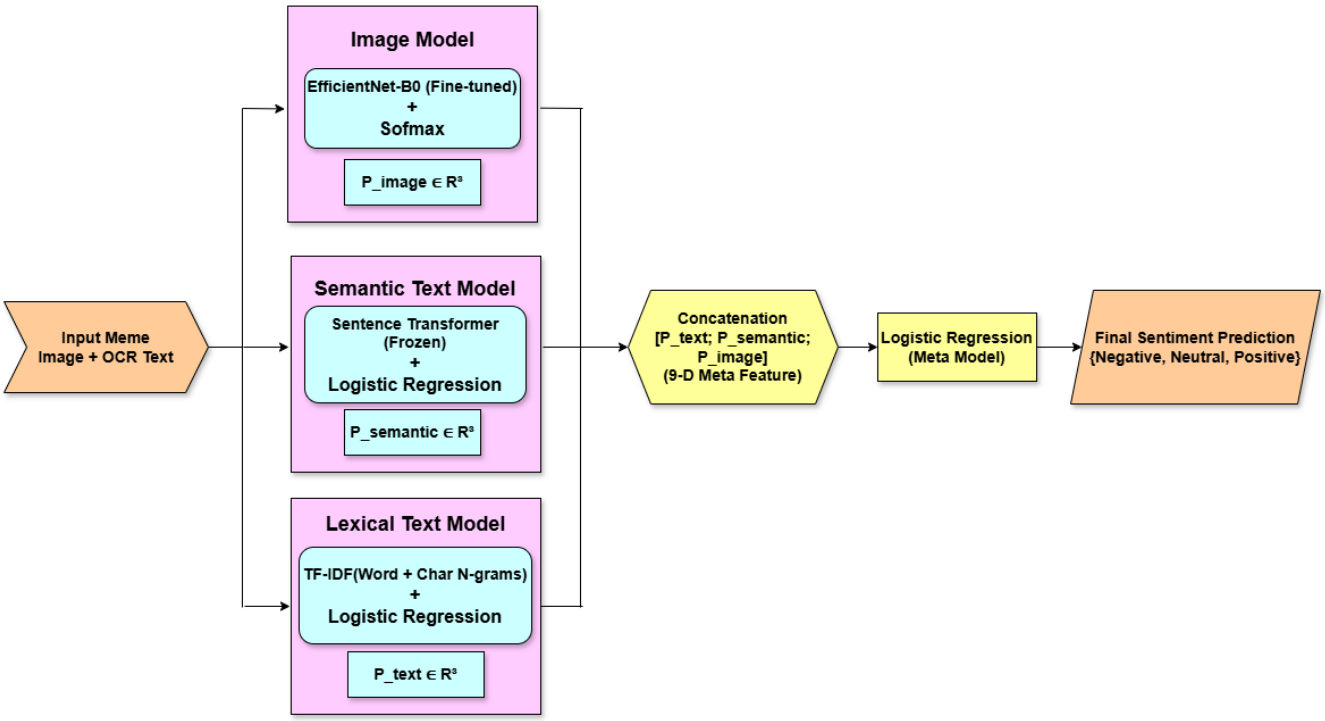

Figure 1: Methodological overview of the proposed method.

4.4. Visual Representation

To capture visual context such as facial expressions and symbolic imagery, we have used EfficientNet-B0 initialized with ImageNet-pretrained weights. The pooled feature vector of dimension 1280 is connected to the classification head.

The model was then fine-tuned using the AdamW optimizer, where the learning rate was set to 1e-4, and the batch size was set to 16. It was then trained for 4 epochs for each fold under stratified 5-fold cross-validation. No early stopping was employed, as the number of epochs for training was small.

The data augmentation techniques employed included resizing the images to 224x224, random horizontal flip, and random rotations ($\pm 10^{\circ}$). For the validation, only image resizing and standardization of the mean and standard deviation of the ImageNet dataset were employed.

The backbone is fine-tuned using the cross-entropy loss function with moderate image augmentation, such as flipping the image horizontally and normalization. Fine-tuning enables adaptation to meme-specific visual patterns.

Prior work, including Bucur et al. (2022) and Roy et al. (2024), have shown that adapting visual backbones improves performance over unimodal approaches. However, we have used the lightweight CNN model rather than the Vision Transformers to maintain stability under limited data.

4.5. Late-Fusion Stacking Strategy

Each unimodal component (lexical text, semantic text, visual model) has produced a probability distribution over the three sentiment classes:

$$P_{\text{text}},\ P_{\text{semantic}},\ P_{\text{image}} \in \mathbb{R}^3$$

These probability vectors were concatenated to form a 9-dimensional meta-feature vector:

$$P_{\text{meta}} = [P_{\text{text}} \parallel P_{\text{semantic}} \parallel P_{\text{image}}]$$

A Logistic Regression meta-classifier was trained on these concatenated probabilities.

Although cross-modal transformers like VisualBERT make direct multimodal reasoning, they can be computationally costly and may suffer from overfitting, especially for small datasets. Inspired by ensemble strategies used in Hateful Memes Challenge in Velioglu and Rose (2020), we have used a late fusion approach for better robustness while maintaining modality-specific inductive biases.

We have used stratified 5-fold cross-validation to generate out-of-fold predictions for stacking, prevent information leakage and improve stability. Finally, predictions from each fold were averaged during the testing phase.

Overall, the multimodal sentiment classification model follows a design that favors modularity, interpretability, and generality.

5. Results and Analysis

5.1. Evaluation Metric

We have used **Macro-F1** to measure the performance of our system, which is also the official metric for the shared task. Macro-F1 assigns equal weight to each class, thus it is suitable for imbalanced sentiment distributions. We have also presented Precision, Recall, and F1 for individual classes to give a comprehensive view of the performance of our system. For all these metrics, we have used stratified 5-fold cross-validation to obtain the results.

5.2. Ablation Study on Modality Combinations

To understand the contribution of each modality in the system, we have conducted an ablation study on the performance of the unimodal models, the two modality combinations, and the complete three modality stacking framework. For this purpose, the Macro-F1 metric is used on the stratified 5-fold cross-validation protocol, and the results are reported in Table 2.

Model Configuration	Macro-F1
TF-IDF (Lexical)	0.3838
SentenceTransformer (Semantic)	0.3774
EfficientNet-B0 (Image)	0.3980
TF-IDF + Image	0.3878
SentenceTransformer + Image	0.3897
TF-IDF + SentenceTransformer	0.3715
TF-IDF + SentenceTransformer + Image	**0.4234**

Table 2: Macro-F1 scores of the unimodal models, the two modality combinations, and the complete three modality stacking framework.

In the case of unimodal models, the image model has the highest score, i.e., 0.3980, which indicates the importance of the visual modality in the sentiment classification of Nepali memes. The best-performing unimodal model i.e., the TF-IDF model, outperforms the semantic SentenceTransformer model, which indicates the importance of the surface-level textual features in the sentiment classification of the memes.

The use of two-modality combinations does not outperform the best unimodal model consistently. Although the TF-IDF + Image pair achieved a competitive result of 0.3878, as did the SentenceTransformer + Image pair with a result of 0.3897, these combinations did not outperform the image model alone. The textual combination of TF-IDF + SentenceTransformer, on the other hand, underperforms with a result of 0.3715, showing redundancy in the use of both textual representations.

On the contrary, the entire three-modality stacking model achieves the best Macro-F1 score of **0.4234**, which reflects a 6.4% improvement over the best unimodal model. This suggests that the use of lexical, semantic, and visual representations offers complementary information, and the use of the three modalities altogether is a prerequisite to improve the sentiment classification of Nepali memes using the multimodal approach.

It is important to note that there is a notable difference between the cross-validation Macro F1 score and the test set Macro F1 score (0.4234 and 0.5045, respectively). This can be explained by the fact that the data may not be evenly distributed in the training and test datasets, and the dataset is relatively small. Since our approach relies on stratified 5-fold cross-validation without post-hoc tuning on the test set, the cross-validation score provides a conservative estimate of model performance, while the higher test score reflects better generalisation on the unseen data.

5.3. Cross-Validation Stability

To measure model's robustness, the Macro-F1 score is provided fold-wise, as shown in Table 3.

Fold	Macro-F1
1	0.419271
2	0.421013
3	0.454797
4	0.407765
5	0.407215

Table 3: Fold-wise Macro-F1 scores

The performance ranges from 0.4072 to 0.4548, with a mean of 0.4220 and a variation range of 0.0476. This shows that the model's performance is quite stable, implying that the stacking model reduces overfitting.

5.4. Per-Class Performance

Table 4 presents class-wise evaluation metrics.

Class	Precision	Recall	F1
Negative	0.4447	0.5777	0.5026
Neutral	0.5375	0.3636	0.4338
Positive	0.3054	0.3684	0.3339

Table 4: Per-class performance of the stacked model

The model performs best on the *Negative* class (F1 = 0.5026). This is because the model's recall is quite high. This shows that the model is good at identifying explicit negative sentiment signals. Performance on the *Neutral* class is moderate with

relatively low recall(F1 = 0.4338). This shows that the model is not good at identifying the class of ambiguous text. The *Positive* class yields the lowest F1 score (0.3339), which shows that the model is not good at distinguishing positive sentiment from neutral or sarcastic content.

5.5. Confusion Matrix Analysis

Figure 2 illustrates the confusion matrix of the final stacked model.

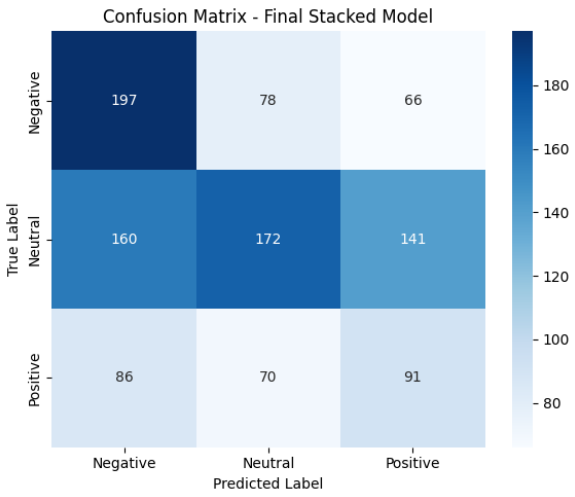

Figure 2: Confusion matrix of the stacked multimodal model.

The confusion matrix shows that there is a significant overlap between the Neutral class and the other two classes. There is a significant number of Neutral instances that have been incorrectly predicted to be Negative and Positive. This shows that the semantic complexity and ambiguity of meme content make sentiment analysis difficult. The Negative class has the highest number of correct predictions, showing that sentiment detection is relatively stable for this class. In contrast, the Positive class is frequently misclassified as Negative, suggesting that subtle positive sentiment and sarcasm in meme content are difficult to detect.

Overall, the proposed late fusion multimodal framework outperforms all the baseline models, and the performance is stable across all cross-validation models. The results confirm that integrating lexical, semantic, and visual representations enhances sentiment classification in low-resource meme settings. However, the moderate Macro F1 score and poor performance of the model on the Positive class show that sentiment understanding in multimodal memes is difficult.

6. Conclusion

In this paper, we have proposed a late fusion multimodal approach for sentiment classification of Nepali text-embedded memes in a low-resource scenario. The proposed method incorporates lexical TF-IDF features, semantic sentence embeddings and fine-tuned image features via a stacking-based meta-classifier to achieve a Macro F1 score of 0.4234 in cross-validation, which surpasses unimodal baselines. On the official test set, the model achieved a higher Macro-F1 score of 0.5045, securing 6th position, which indicates strong generalisation despite the limited training data. The image-based model exhibits the best performance individually, indicating the importance of image content in meme sentiment classification. The stacking-based fusion effectively combines modality-specific features while maintaining consistent performance across different folds in cross-validation. The overall performance is moderate, while there are challenges in classifying the Positive class. The proposed method can be a starting point to explore better cross-modality interaction mechanisms to improve sentiment classification. The overall study can serve as a reproducible and modular baseline for multimodal sentiment classification in low-resource meme datasets.

Overall, this study provides a reproducible and modular baseline for multimodal sentiment classification in low-resource meme datasets.

7. Limitations

Although the suggested late fusion multimodal model presents competitive results, there are a number of limitations that need to be noted. Firstly, the dataset size is not large, as there are only 1,061 training samples. This makes it difficult for deep models to fully grasp the complex relationships between the modalities. The model also seems to have difficulty in recognizing Positive sentiment; this could be due to difficulties in recognizing sarcasm or irony.

Secondly, the suggested model relies on text features obtained through OCR from the image dataset, which may not capture culturally specific features that may be part of the meme. Finally, the class imbalance that may affect the detection of certain classes. In future work, there may be a number of directions that can be explored, including data augmentation, class imbalance, and cross-modal interaction.

8. Ethics Statement

This study undertook a multimodal sentiment analysis of Nepali memes and this was conducted within

a shared-task framework. The dataset was made available by the organizers. No personal data was collected for this study.

It is to be noted that, although sentiment analysis tools are helpful for understanding content and promoting digital well-being, automated models may also contain biases, which are usually reflected by the data on which these models are trained. These biases may result in misclassifications, particularly for ambiguous and sarcastic memes, which are common in social media.

Thus, we would like to emphasize that, multimodal sentiment models should be used responsibly and human intervention should also be considered to ensure fairness and context-awareness, particularly for such models that may contain biases.

9. 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.

Ana-Maria Bucur, Adrian Cosma, and Ioan-Bogdan Iordache. 2022. Blue at memotion 2.0 2022: You have my image, my text and my transformer.

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

A. Guleria, K. Varshney, G. Pahwa, et al. 2025. Multimodal sentiment analysis of english and hinglish memes. *Multimedia Tools and Applications*, 84:15331–15356.

Md Hasan, Nusratul Jannat, Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. CUET-NLP@DravidianLangTech-ACL2022: Investigating deep learning techniques to detect multimodal troll memes. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 170–176, Dublin, Ireland. Association for Computational Linguistics.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. MemoSen: A multimodal dataset for sentiment analysis of memes. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1542–1554, Marseille, France. European Language Resources Association.

Rishav Jha, Mohit Ranjan Panda, Shubham K C, and Aaditya Dahal. 2025. Deep learning architectures for multimodal sentiment analysis. In *2025 International Conference on Intelligent and Cloud Computing (ICoICC)*, pages 1–6.

Nicola Messina, Fabrizio Falchi, Claudio Gennaro, and Giuseppe Amato. 2021. AIMH at SemEval-2021 task 6: Multimodal classification using an ensemble of transformer models. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1020–1026, Online. Association for Computational Linguistics.

Aishwaria Roy, Hasan Murad, and Ayman Iktidar. 2024. A multimodal approach to bangla cyberbullying meme detection in social media. In *2024 27th International Conference on Computer and Information Technology (ICCIT)*, pages 1886–1891.

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)*.

Kogilavani Shanmugavadivel, Malliga Subramanian, Mohamed Arsath H, Ramya K, and Ragav R. 2025. TEAM_STRIKERS@DravidianLangTech2025: Misogyny meme detection in Tamil using multimodal deep learning. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 619–623, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Mayukh Sharma, Ilanthenral Kandasamy, and W.b. Vasantha. 2020. Memebusters at SemEval-2020 task 8: Feature fusion model for sentiment analysis on memes using transfer learning. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1163–1171, Barcelona (online). International Committee for Computational Linguistics.

Shivam Sharma, Ramaneswaran S, Md. Shad Akhtar, and Tanmoy Chakraborty. 2024. Emotion-aware multimodal fusion for meme emotion detection. *IEEE Transactions on Affective Computing*, 15(3):1800–1811.

Steve Durairaj Swamy, Shubham Laddha, Basil Abdussalam, Debayan Datta, and Anupam Jamatia. 2020. NIT-agartala-NLP-team at SemEval-2020 task 8: Building multimodal classifiers to tackle Internet humor. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1179–1189, Barcelona (online). International Committee for Computational Linguistics.

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.

Georgi Vankov, Dimitar Dimitrov, Ivan Koychev, and Preslav Nakov. 2025. Multimodal sentiment analysis: Recognizing sentiment in memes. In *Artificial Intelligence: Methodology, Systems, and Applications*, pages 1–11, Cham. Springer Nature Switzerland.

Riza Velioglu and Jewgeni Rose. 2020. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge.