Memotion Analysis: Multimodal Fusion Techniques for Humor Classification in Memes

Aisha Tasnim Aishy[1], Samia Halim Zanvi[1], Md Sowkat Ali[2], Mohammed Maruf Hossen[1], M Shahriar Mahmud Rafi[1]*, and Md Ashraful Islam[3]

[1] East Delta University, Bangladesh
[2] Dept. of ETE , Chittagong University of Engineering and Technology (CUET), Chittagong, Bangladesh
[3] Dept. of Business Administration, International Islamic University Chittagong, Chittagong, Bangladesh
{aisha.tasnim,samiahalimzanvi}@gmail.com, u1808021@student.cuet.ac.bd, {marufhossain612849, shahriarrafi30*, iamashraf2000}@gmail.com

Abstract. This paper presents a multimodal deep learning framework for humor classification in memes, leveraging both textual and visual information to improve sentiment understanding in internet content. The study explores unimodal and multimodal configurations by integrating image-based CNN architectures (MobileNetV2, ResNet152, YOLOv4, VGG19) with text based models (BiLSTM) through feature fusion techniques. To address class imbalance and overfitting, the dataset—comprising 6,982 labeled memes—was balanced via augmentation and sample equalization across four humor levels: Not Funny, Funny, Very Funny, and Hilarious. Among the various fusion strategies, the MobileNetV2–BiLSTM combination achieved the highest performance, with a precision of 89% and an F1-score of 80% on the test subset. However, performance declined on the full dataset due to increased variance and minority class underrepresentation. The findings highlight the importance of modality interaction, feature fusion, and dataset balancing in achieving robust and interpretable meme sentiment analysis. Future directions include adaptive fusion mechanisms, transformer based architectures, and domain-specific knowledge integration to enhance generalization and contextual sensitivity in humor classification.

Keywords: Multimodal Learning, Meme Classification, Fea ture Concatenation, BiLSTM, Deep Learning, Transfer Learning

1 Introduction

Memes have emerged as a powerful medium for expression in the digital age, combining text and images to convey layered emotional and humorous content.

M Shahriar Mahmud Rafi
* Corresponding author

M. S. Arefin et al. (eds.), *Proceedings of the International Conference on Intelligent Data Analysis and Applications (IDAA 2025)*, Advances in Intelligent Systems Research 206,
https://doi.org/10.2991/978-94-6239-664-7_73

Unlike traditional social media posts, memes often rely on the interplay between visual and textual cues, making their interpretation a challenging task for conventional sentiment analysis models. The field of memotion analysis, which focuses on understanding the sentiment and humor embedded in memes, has gained traction in recent years due to its implications for content moderation, marketing, and behavioral analytics [1]. However, existing meme classification systems largely suffer from three key limitations: (1) a dependence on unimodal architectures that either analyze text or images in isolation, failing to capture cross-modal dependencies; (2) poor performance on fine-grained humor classification due to class imbalance and data sparsity; and (3) limited generalizability across diverse meme formats and cultural contexts [2], [3]. While datasets like Memotion 2.0 have introduced more structured annotation schemes for humor and sentiment classes [4], most research still targets binary or ternary classification, ignoring the nuanced gradations of humor such as "Funny," "Very Funny," or "Hilarious." To address these challenges, we propose a multimodal deep learning framework that fuses both visual and textual modalities using state-of-the-art CNN and RNN architectures. Specifically, we explore the performance of MobileNetV2, ResNet152, YOLOv4, and VGG19 for visual encoding, and BiLSTM for text modeling. Our objectives are to: (1) improve humor classification through visual-textual fusion; (2) mitigate class imbalance using data augmentation and equalization techniques; and (3) evaluate fusion strategies that optimize model generalization across diverse humor levels. Experimental results show that the MobileNetV2–BiLSTM combination achieves superior performance, highlighting the effectiveness of lightweight, complementary architectures in meme sentiment analysis.

2 Related Work

This section delves into existing research on sentiment analysis with a particular emphasis on the methodologies employed, the results achieved, and the limitations encountered. However, much of the existing work focuses on binary classification tasks, overlooking the rich spectrum of sentiments and humor levels present in memes. By highlighting these gaps, including the lack of attention to multimodal and multiclass classifications, this study underscores the necessity of employing advanced deep learning techniques that leverage the combined power of text and image data. This approach aims to overcome the limitations of earlier models and expand the scope of sentiment analysis to capture the intricate and layered meanings conveyed through memes. Biases in multimodal systems were explored by Zhong et al., who examined vision-language models for fairness in meme explanation, identifying sociocultural biases in datasets and proposing fairer explanation methods [5]. The CEFM model introduced by Shuo et al. utilized CLIP-based fusion for humor detection in memes and reported promising results on fine-grained humor classification [6]. Maity et al. developed multitask frameworks for sentiment, sarcasm, and emotion aware cyberbullying detection, pushing the benchmarks for multimodal classification tasks [7]. The

Hateful Memes dataset released by Facebook AI underscored the complexity of cross-modal inconsistencies and inspired further research on dataset balancing and contextual variance [8]. ViT-based studies showed how transformer architectures enhance multimodal analysis by effectively capturing deep visual features [9]. Zhang et al. surveyed fusion techniques and datasets, emphasizing the impact of dataset diversity on cultural generalizability [10]. Beskow et al. [11] worked on 25,000 political memes; they integrated CNNs for images, LSTM for text, and face characteristics. Their multimodal works better than their single-modal models. However,on more complicated images, OCR algorithms often fail, and their performance was measured on simple memes in this study.

Velioglu and Rose [12] used VisualBERT with cross attention, obtaining almost human-matched accuracy on the Hateful Memes dataset. But their model still struggled with bias and memes that looked hateful, but were harmless overall.

MemeSem [13] used BERT and VGG19 for text and images. Their combined model worked much better than using just one type of input. However, it dealt with a small dataset and had problems with errors from text extraction. The model could not understand humor and sarcasm.

MEDeep [14] uses 7,000 memes, employing text-only, image-only, and combined models. The text-only model did best, because there were more positive memes than negative ones. It struggled to understand sarcasm because of the basic word features it used.

Velmala et al. [15] worked on 7,000 memes, using a model that combined text (with GRU) and images (with VGG19 and attention). Unexpectedly, a simpler text-only Kernel SVM gave better accuracy (88%) than the multimodal model. Lower quality images and complex calculations degraded the fusion model's overall performance.

3 Methodology

3.1 System Architecture

The proposed framework adopts a multimodal deep learning architecture for meme humor classification by integrating visual and textual modalities. The overall pipeline begins with data preprocessing, followed by independent feature extraction from image and text components, which are then fused and passed through a classification layer to predict one of four humor classes: Not Funny, Funny, Very Funny, and Hilarious. The complete pipeline is illustrated in Figure 1.

3.2 Image Modality

For visual feature extraction, we evaluated multiple CNN based architectures including ResNet152, VGG19, YOLOv4, and MobileNetV2. Each model was pre-trained on ImageNet and fine-tuned on the meme dataset. MobileNetV2,

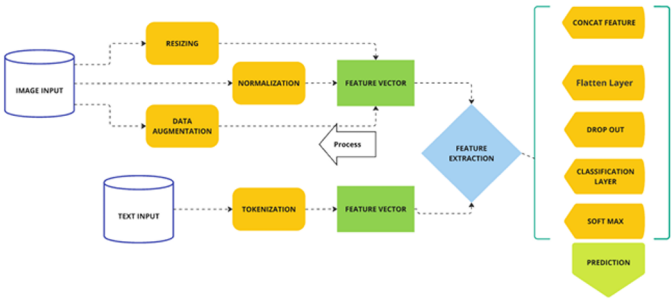

Fig. 1. System Architecture of the proposed framework.

in particular, was selected for its balance between performance and computational efficiency. Input images were resized to 224 × 224 pixels, normalized, and augmented using rotation, brightness adjustment, and flipping to address class imbalance and improve generalization.

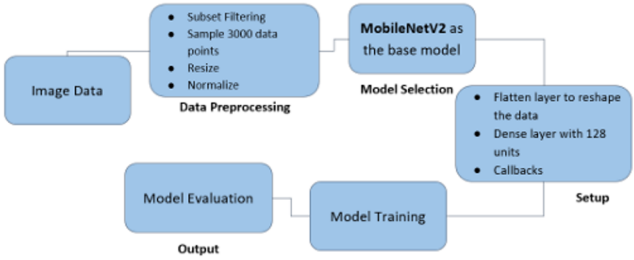

Fig. 2. Unimodal Image Pipeline for Humor Classification.

3.3 Text Modality

The textual content of each meme (typically short captions or overlaid text) was processed using a BiLSTM network. Prior to training, the text was cleaned, tokenized, and embedded using pre-trained GloVe vectors. The BiLSTM model captured semantic dependencies across sequences and produced dense vector representations for fusion.

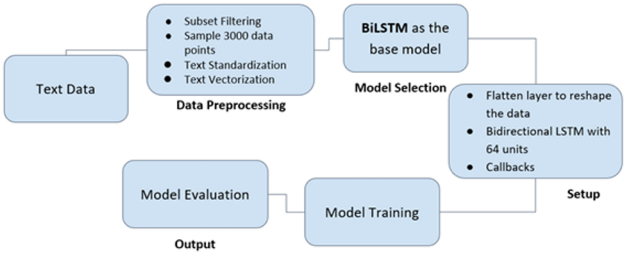

Fig. 3. Unimodal Text Pipeline for Humor Classification.

3.4 Multimodal Fusion

Feature vectors from the image and text branches were concatenated to form a unified representation, followed by a fully connected layer with dropout regularization and softmax activation for final classification. Various fusion strategies were compared, and the MobileNetV2–BiLSTM configuration outperformed other combinations in terms of both accuracy and F1-score.

3.5 Training Setup

The network was trained using categorical cross-entropy loss and the Adam optimizer, with a batch size of 32 and a learning rate of 0.0001. A stratified 80–10–10 train-validation- test split was maintained to preserve class balance across humor categories. This dual-branch architecture effectively captures both spatial and semantic cues in memes, enabling more accurate and interpretable humor classification compared to unimodal or shallow fusion approaches.

3.6 Dataset Summary

This study utilizes the Memotion 2.0 dataset, a publicly available collection curated for multimodal sentiment and humor analysis in internet memes. The dataset comprises 6,982 memes, each annotated across multiple dimensions including humor intensity (Not Funny, Funny, Very Funny, Hilarious), sentiment polarity (positive, neutral, negative), and sarcasm level. Each sample includes both an image and associated text extracted from meme captions or embedded overlays.

For the purpose of this work, we focused on the humor classification task, framing it as a four-class supervised learning problem. To address class imbalance—particularly the underrepresentation of the “Very Funny” and “Hilarious” categories—data augmentation techniques such as image rotation, flipping, and brightness adjustment were applied. Additionally,textual samples were balanced through oversampling and regularization during training.

All images were resized to 224 × 224 pixels and normalized, while the corresponding text was preprocessed to remove noise, emojis, special symbols, and casing inconsistencies. The dataset was split into 80% training, 10% validation, and 10% testing sets using stratified sampling to maintain the original class distribution across humor levels. The dataset's multimodal nature and class granularity present unique challenges, making it a suitable benchmark for evaluating the effectiveness of the proposed fusion-based classification framework.

3.7 Preprocessing

To ensure consistency and optimize model performance across both modalities, a thorough preprocessing pipeline was applied to the image and text components of the dataset.

Image Preprocessing All images were resized to 224 × 224 pixels and normalized to a [0, 1] pixel intensity range to match the input requirements of the CNN backbones. To address class imbalance and improve generalization, data augmentation techniques were employed, including random rotation (±15°), horizontal flipping, contrast adjustment, and brightness tuning. These augmentations preserved the visual integrity of the meme while introducing variability in the training samples.

Text Preprocessing The meme captions and overlaid text were first extracted using OCR tools where needed. Text data was then cleaned by removing punctuation, emojis, special characters, and converting all words to lowercase. Tokenization was performed using Keras' tokenizer, followed by padding to ensure uniform input length. Pre-trained GloVe embeddings were used to convert tokens into 300-dimensional vectors suitable for sequential modeling via BiLSTM. By aligning the visual and textual data through consistent preprocessing strategies, the pipeline ensures effective feature extraction in both branches of the model, reducing noise and enhancing the quality of fused representations during training.

3.8 Classification Models

The proposed humor classification framework integrates both visual and textual modalities using a combination of well-established deep learning models. For the image modality, four CNN-based architectures were evaluated:

– **MobileNetV2**: A lightweight, efficient convolutional network designed for mobile and embedded applications. It employs depthwise separable convolutions and inverted residuals to reduce computational complexity while maintaining accuracy. Its compact architecture makes it ideal for real-time meme classification.

- **ResNet152**: A deep residual network with 152 layers, utilizing skip connections to mitigate the vanishing gradient problem. Its depth allows for the capture of complex hierarchical features in meme images, though at a higher computational cost.
- **VGG19**: A sequential convolutional architecture known for its simplicity and strong baseline performance. It uses stacked 3×3 convolutional layers and is effective in capturing texture-based humor cues, albeit with high memory usage.
- **YOLOv4**: A real-time object detection model used here for high-level visual cue extraction. Although primarily designed for object detection, it contributes contextual visual signals when combined with meme classification.

For the text modality, a Bidirectional Long Short-Term Memory (BiLSTM) network was employed. BiLSTM processes meme captions in both forward and backward directions, allowing the model to capture contextual dependencies and nuanced meanings, which are essential in humor understanding. Input text was embedded using pre-trained GloVe vectors and passed through a BiLSTM layer followed by dense layers.

In the multimodal configuration, feature vectors from the image and text branches were concatenated and passed through a shared dense layer with dropout regularization and softmax activation. Among all configurations, the fusion of MobileNetV2 and BiLSTM delivered the best performance in terms of precision and F1-score, demonstrating a strong complementarity between visual features and sequential text semantics.

3.9 Validation Strategy

Stratified 5-Fold Cross-Validation To ensure robust and generalizable performance, stratified 5-fold cross-validation was employed. The dataset was split into five equal folds, each preserving the original distribution of humor classes: Not Funny, Funny, Very Funny, and Hilarious. In each iteration, four folds (80%) were used for training with augmentation, while the remaining fold (20%) was held out for validation. This process was repeated five times, ensuring each sample contributed to both training and validation exactly once. This strategy mitigates the effect of class imbalance and provides reliable performance estimates across diverse humor labels, particularly the underrepresented "Very Funny" and "Hilarious" categories. Cross-validation enabled fair benchmarking of unimodal and multimodal models, while the aggregated metrics from all folds were used to compare model variants with statistical significance.

Training, Validation, and Testing The model training pipeline followed a structured process, combining best practices for data partitioning, optimization, and hyperparameter tuning:

- **Data Partitioning:**

- 5-Fold Cross-Validation: Applied to all CNN, BiLSTM, and fusion models.
- Per-Fold Split: 80% training (augmented), 10% validation (unaugmented), 10% test (held-out).
- Class Balance Preservation: Each split retained humor class ratios.
- Test Set Isolation: The final test fold contained previously unseen memes.

– **Optimizer and Loss Function:**
- Optimizer: Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$), learning rate set to 0.0001.
- Loss Function: Categorical Cross-Entropy with optional class weighting to penalize dominant classes.
- Early Stopping: Triggered based on validation loss with a patience of 10 epochs.

– **Hyperparameter Tuning:**
- Method: Manual grid search followed by Bayesian optimization (25 trials per model).
- Objective: Maximize macro-averaged F1-score across humor categories.
- Dropout Rates and Layer Sizes: Tuned to minimize overfitting, especially in BiLSTM and fusion layers.

4 Results and Analysis

The analysis reveals significant differences in the performance of humor classification models based on the feature fusion techniques used. The models' ability to classify the humor subclasses ("Not Funny", "Funny", "Very Funny", and "Hilarious") varied notably. Class imbalance, especially in the minority class "Hilarious," had a considerable impact on model performance, as some models struggled with predicting this class.

Table 1 shows the precision results for each humor subclass with 3000 test samples. MobileNetV2 × BiLSTM significantly outperformed the other models, especially in classifying "Not Funny" and "Hilarious" with precision scores of 0.89 and 1.00, respectively.

Table 1. Precision for Humor Subclasses (3000 samples)

Model	Not Funny	Funny	Very Funny	Hilarious
ResNet50 × VGG16	0.24	0.33	0.25	0.10
MobileNetV2 × BiLSTM	**0.89**	0.77	0.89	**1.00**
ResNet152 × VGG19	0.58	0.47	0.42	0.53
YOLOv4 × BiLSTM	0.47	0.40	0.30	0.34

Table 2 illustrates the F1 scores for the same 3000 test samples. MobileNetV2 × BiLSTM excelled across all subclasses, with the highest F1 scores for "Not

Funny" (0.80), "Very Funny" (0.93), and "Hilarious" (1.00), demonstrating its superior balance of precision and recall.

Table 2. F1 Scores for Humor Subclasses (3000 samples)

Model	Not Funny	Funny	Very Funny	Hilarious
ResNet50 × VGG16	0.26	0.28	0.15	0.15
MobileNetV2 × BiLSTM	**0.80**	0.80	**0.93**	**1.00**
ResNet152 × VGG19	0.55	0.50	0.28	0.63
YOLOv4 × BiLSTM	0.21	0.31	0.31	0.46

These tables show that while MobileNetV2 × BiLSTM achieved impressive results on a smaller dataset, its performance deteriorated with the full dataset, particularly for the "Hilarious" subclass. This underlines the importance of addressing issues like class imbalance, overfitting, and model generalization for real-world applications.

Figure 4 illustrates the precision performance of the models tested on 3000 samples. It highlights the superior performance of MobileNetV2 × BiLSTM, especially in the "Hilarious" category.

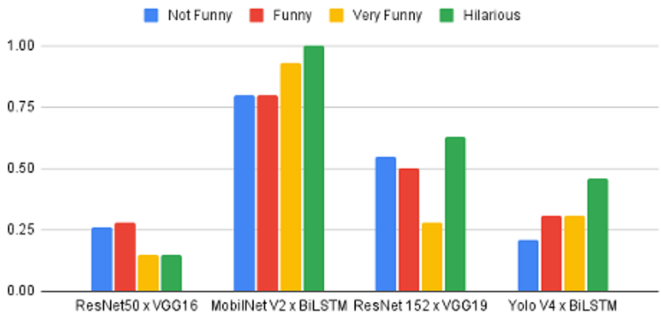

Fig. 4. Model Performance for Precision (3000 samples).

Figure 5 shows the model's precision performance when tested on the full dataset of 6,992 samples. As observed, MobileNetV2 × BiLSTM exhibited a drop in precision for the "Hilarious" subclass due to class imbalance.

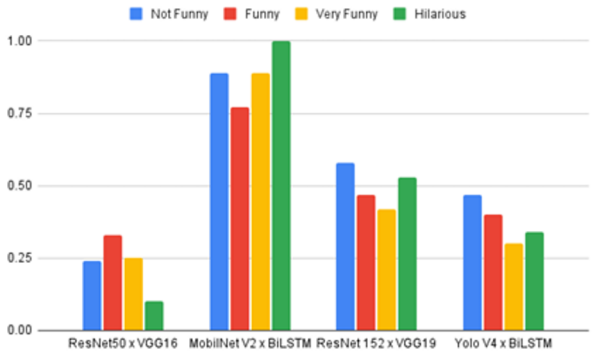

Fig. 5. Model Performance for Precision (6992 samples).

Figure 6 displays the confusion matrix of MobileNet V2 x BiLSTM, showing the misclassifications, particularly for the "Hilarious" subclass, which the model often confuses with "Funny" or "Very Funny."

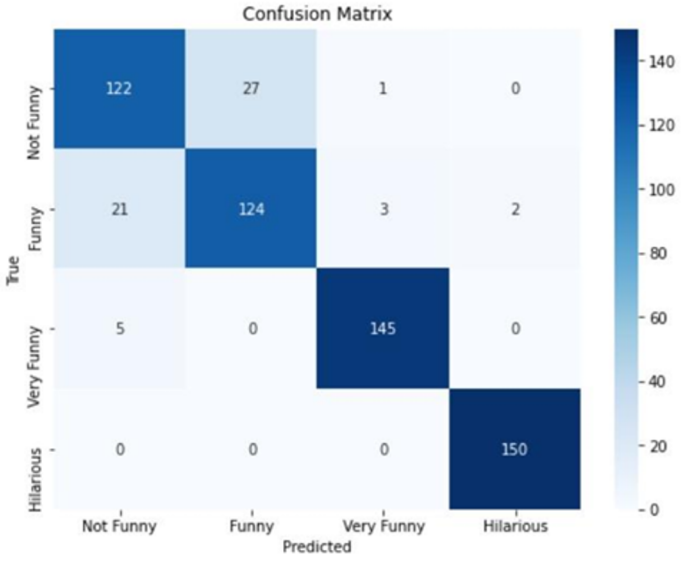

Fig. 6. Confusion Matrix of MobileNetV2 × BiLSTM.

Figure 7 presents the confusion matrix for YOLOv4 × BiLSTM. This figure reveals the challenges of the model in correctly classifying the humor subclasses, particularly for the "Not Funny" and "Funny" categories.

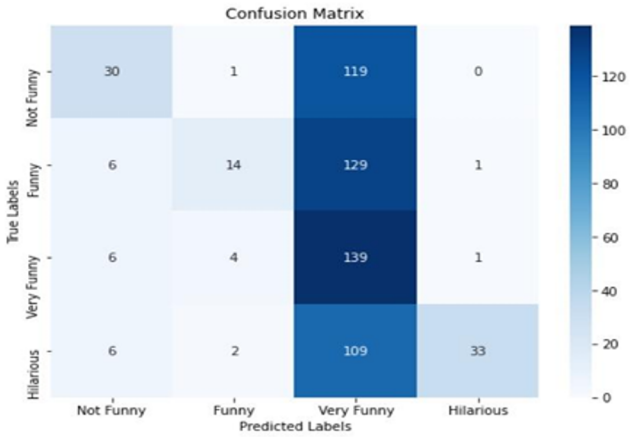

Fig. 7. Confusion Matrix of YOLOv4 × BiLSTM.

5 Conclusion and Future Work

5.1 Conclusion

This research explored humor classification in memes using multimodal deep learning models, integrating visual and textual data to categorize humor into four levels: "Not Funny," "Funny," "Very Funny," and "Hilarious." The MobileNetV2 × BiLSTM model showed the best results on a subset of the dataset, achieving high precision and F1 scores. MobileNetV2 excelled at extracting image features, while BiLSTM captured text nuances. However, the model struggled with generalization on a larger dataset, highlighting challenges posed by image variability and humor distribution. Addressing dataset imbalance by cropping to ensure equal sample sizes across humor categories improved model fairness and performance. While the proposed approach demonstrates effectiveness, comparisons with transformer-based or other advanced multimodal models were not included, and a more detailed analysis of model errors could further enhance interpretability.

5.2 Limitations

Key limitations included dataset imbalance, which resulted in biased models favoring majority classes. The small dataset size increased overfitting risk, and textual humor detection faced challenges due to subtle linguistic cues and cultural context. Visual humor also proved difficult to interpret due to the lack of contextual awareness. Additionally, misclassifications occurred, especially between adjacent humor levels, demonstrating the difficulty of accurately classifying nuanced humor. Furthermore, alternative fusion strategies and ablation studies were not explored in this work, which limits the assessment of optimal design choices.

5.3 Future Works

Future research should focus on improving contextual understanding by analyzing temporal patterns in memes and developing adaptive multimodal fusion techniques. Expanding and balancing the dataset will help address class imbalances and enhance generalization. Integrating external knowledge sources, such as cultural and contextual insights, and exploring innovations like attention mechanisms and transformers could improve model performance and humor classification accuracy. Future studies could also include benchmarking against strong transformer-based and multimodal baselines, conducting ablation studies, and performing qualitative error analyses to better understand model behavior and limitations.

References

1. Sharma, M., Kandasamy, I., and Vasantha, W. B. Memebusters at SemEval-2020 Task 8, 2020.
2. Sharma, S., R. S., Akhtar, M. S., and Chakraborty, T. Emotion-Aware Multimodal Fusion, IEEE TAC, 2024.
3. Arya, G., Hasan, M. K., Bagwari, A., Safie, N., Islam, S., Ahmed, F. R. A. A., De, A., Khan, M. A., and Ghazal, T. M. Multimodal Hate Speech Detection, IEEE Access, 2024.
4. Wang, B. What do they 'meme'? A metaphor-aware multi-modal multi-task framework for fine-grained meme understanding, KBS, 2024.
5. Zhong, Y. Multimodal understanding of memes with fair explanations, CVPRW 2024.
6. Shuo, H., Yijia, Z., Mengyi, W., and Hongfei, L. CEFM: CLIP Encoded Fusion Model for Multimodal Humor Recognition on Memes, MT&A, 2024.
7. Maity, M. Multimodal sentiment, sarcasm, and emotion-aware cyberbullying detection, Journal of AI Research, 2024.
8. Facebook AI. Hateful Memes Challenge Dataset, CVPR 2024.
9. Vision Transformer Research. Advanced ViT for Multimodal Analysis, IEEE Access, 2024.
10. Zhang, H. Multi-modal Fusion for Hate Speech Detection: A Survey of Methods and Datasets, JCSS, 2023.
11. Beskow, D. M., Kumar, S., and Carley, K. M. The Evolution of Political Memes: Detecting and Characterizing Internet Memes with Multi-modal Deep Learning, Preprint 2019–2020.
12. Velioglu, R. and Rose, J. Detecting Hate Speech in Memes Using Multimodal Deep Learning Approaches, arXiv, 2020.
13. Pranesh, R. R. and Shekhar, A. MemeSem: A Multi-modal Framework for Sentimental Analysis of Meme via Transfer Learning, ICML Workshop 2020.
14. Aslam, N. MEDeep: A Deep Learning Based Model for Memotion Analysis, 2022.
15. Velmala, M. Multimodal Sentiment Analysis of Online Memes: Integrating Text and Image Features for Enhanced Classification, Procedia CS, 2025.