Multimodal Deep Learning for Detection of Hate, Humor, and Stance in Social Discourse on Marginalized Communities

Durgesh Verma
IITRAM, Ahmedabad
Gujarat, India
durgesh.verma.24co@iitram.ac.in

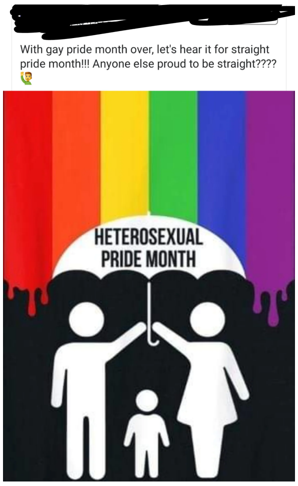
Abhinav Kumar
CSED, MNNIT Allahabad
Prayagraj, India
abhik@mnnit.ac.in

Abstract

Internet memes serve as powerful vehicles of expression across platforms like Instagram, Twitter, and WhatsApp. However, they often carry implicit messages such as humor, sarcasm, or offense especially in the context of marginalized communities. Understanding such intent is crucial for effective moderation and content filtering. This paper introduces a deep learning-based multimodal framework developed for the CASE 2025 Shared Task on detecting hate, humor, and stance in memes related to marginalized movements. The study explores three architectures combining textual models (BERT, XLM-RoBERTa) with visual encoders (ViT, CLIP), enhanced through cross-modal attention and Transformer-based fusion. Evaluated on four subtasks, the models effectively classify meme content—such as satire and offense—demonstrating the value of attention-driven multimodal integration in interpreting nuanced social media expressions.

1 Introduction

Memes have emerged as a dominant medium of communication in the digital age, enabling users to express emotions, opinions, and social commentary in humorous yet impactful ways. Their wide dissemination on platforms such as Twitter, Instagram, and WhatsApp makes them not only vehicles of entertainment but also instruments of cultural and ideological expression. Despite their seemingly innocuous appearance, memes can carry coded language, sarcasm, and implicit ideologies that may reinforce hate (Parihar et al., 2021; Roy and Kumar, 2025; Swain et al., 2022; Kumar et al., 2021), misinformation, or discrimination (Zannettou et al., 2018). Their interpretative flexibility often depends on the viewer's cultural background, context, and personal values (Kiela et al., 2020), making automatic intent recognition particularly challenging.



[redacted]

With gay pride month over, let's hear it for straight pride month!!! Anyone else proud to be straight????

Figure 1: A meme promoting 'Heterosexual Pride Month' — raising concerns about LGBTQ+ exclusion.

What makes memes powerful is also what complicates their analysis: they integrate both visual and textual elements, with meaning frequently emerging from the interaction between the two modalities. A single meme can appear humorous to some while being deeply offensive to others. This inherent ambiguity necessitates sophisticated approaches to computational analysis that can reason across modalities and cultural contexts.

Figure 1 illustrates the importance of multimodal reasoning. This meme, sourced from the CASE 2025 shared task (Thapa et al., 2025a; Hürriyetoğlu et al., 2025), visually depicts a heterosexual family shielding their child from colored rain. While it may appear neutral or protective at first glance, the colored rain can be interpreted as representing LGBTQ+ pride, implying an exclusionary and discriminatory undertone.

Interpretation 1: Neutral Perspective: At first glance, the meme may appear to convey a positive or protective sentiment—parents shielding their child from colorful rain. This can be interpreted as

Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025, pages 98–106, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-012

a metaphor for responsible parenting, without any harmful connotation.

Interpretation 2: Critical Perspective: Upon closer examination, the rainbow-colored rain suggests a symbolic representation of LGBTQ+ identity. The umbrella labeled "Heterosexual Pride Month" implies protection from LGBTQ+ influence, thus reinforcing anti-LGBTQ+ sentiment and promoting a harmful ideological stance.

The latest developments in the impressive deep learning, particularly in the field of multimodal learning, have allowed extraction and reasoning of both textual and visual features. The alignment of vision and language, which is vital in the understanding of the layered semantics of memes, has had an impressive result on the architectures, including CLIP (Radford et al., 2021) and BLIP (Li et al., 2022). Particular to meme analysis, the Hateful Memes Challenge (Thapa et al., 2024) and Memotion Analysis tasks (Sharma et al., 2020) have inspired new studies on multimodal hate (Thapa et al., 2023; Bhandari et al., 2023) and sentiment analysis.

It has been observed that unimodal learning systems cannot capture subtle contextual data; hence, to address the situation, we present a multimodal deep learning framework to integrate visual and textual information in a multi-modal environment by means of the fusion strategy of attention. The given approach can identify not only direct but also expressive forms of hate, such as sarcasm and positions of ideology. Our model follows a similar pattern but uses pre-trained architectures (XLM-R on the text data, CLIP model on the visual information) and colleges a Transformer-based fusion module to enable more robust performance with better interpretability on multiple downstream standardized data sets.

The rest of the sections are organized as follows: Section 2 discusses related work for memes identification, Section 3 discusses proposed methodology in detail, The outcome of the proposed model is listed in Section 4 and Section 5 concludes the paper.

2 Related Work

Detecting harmful or misleading memes presents a significant challenge due to their inherently multimodal structure often requiring nuanced understanding of visual cues and embedded text. Over the years, multiple approaches have been developed to address this challenge, ranging from unimodal to state-of-the-art multimodal architectures.

2.1 Text and Vision Models for Content Moderation

Early approaches primarily focused on either the textual or the visual component of memes. Models like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) were widely used for text analysis, while CNN-based models such as ResNet (He et al., 2016) and DenseNet (Huang et al., 2017) handled image classification. However, these unimodal methods often failed to capture the interaction between image and text, a critical element in meme understanding.

2.2 Multimodal Detection and the Facebook Hateful Memes Challenge

The Facebook Hateful Memes Challenge (Kiela et al., 2020) emphasized the need for multimodal solutions by presenting memes where neither text nor image alone conveyed hate. Transformer-based models such as ViLBERT (Lu et al., 2019) offered early solutions for joint vision-language learning.

The introduction of CLIP (Radford et al., 2021) further advanced this field by aligning visual and textual representations in a shared embedding space. Leveraging this, Hate-CLIPper (Kumar and Nandakumar, 2022) and MemeCLIP (Shah et al., 2024) demonstrated robust performance for hateful meme detection and multi-label classification tasks such as humor, stance, and hate.

2.3 Recent Advances in Meme Understanding

Recent studies have further refined multimodal fusion strategies. Align-before-Attend (Hossain et al., 2024) aligns image and text features prior to fusion to improve hate detection performance, particularly on multilingual datasets. Evolver (Huang et al., 2024) applies prompt-based chain-of-evolution reasoning, enabling the model to use historical meme context for interpreting intent.

LMM-RGCL (Mei et al., 2025) introduces a two-stage contrastive learning approach for fine-tuning large multimodal models and achieves state-of-the-art results across six meme datasets. Lin et al. (Lin et al., 2024) propose an explainable debate framework between LLMs (Thapa et al., 2025b) for modeling conflicting viewpoints within memes. M3Hop-CoT (Kumari et al., 2024) uses a multimodal chain-of-thought strategy to enhance misogynous meme detection performance, especially in datasets like MAMI.

In multilingual settings, Chauhan and Kumar (Chauhan and Kumar, 2025) employ XLM-RoBERTa with ViT and BiLSTM-attention for detecting misogyny in Tamil and Malayalam memes. GuardHarMem (El-amrany et al., 2025) incorporates caption generation with fusion-based detection for improved interpretability and performance (F1 ≈ 0.91).

2.4 Research Gap

Despite recent advances in multimodal learning, many existing approaches still rely heavily on task-specific architectures, handcrafted feature engineering, or late fusion strategies that treat textual and visual modalities in isolation until the final stage. These limitations restrict the models' ability to capture fine-grained interactions between modalities and often reduce their generalizability across tasks.

3 Methodology

Each meme sample contains both image and OCR text. All subtasks are multi-class or binary classification problems, requiring both modalities for accurate prediction (see Table 1 for task description). The additional information on the dataset can be found in (Thapa et al., 2025a).

This section presents two multimodal architectures designed for the Shared Task on Multimodal Hate, Humor, and Stance Detection in Marginalized Movements @CASE 2025 (Thapa et al., 2025a; Hürriyetoğlu et al., 2025). Both architectures process meme images and their OCR-extracted text, aiming to predict four semantic properties: hatefulness, targeted group, stance, and humor. The overall task is framed as a multi-task learning problem. The flow diagram for the proposed model can be seen in Figure 2.

Our work introduces a unified transformer-based architecture that leverages early fusion of multimodal features through cross-modal attention. Specifically, we explore three distinct combinations of pre-trained language and vision encoders:

- **XLM-RoBERTa + CLIP**: We concatenate 768-dimensional textual embeddings from XLM-RoBERTa with 512-dimensional text and 512-dimensional image embeddings from CLIP, forming a comprehensive 1792-dimensional multimodal representation.
- **BERT + ViT**: This configuration fuses 768-dimensional text embeddings from BERT (Devlin et al., 2019) with 768-dimensional image embeddings from the Vision Transformer (ViT), resulting in a 1536-dimensional joint feature space.
- **XLM-RoBERTa + ViT**: Here, both text and image features are of 768 dimensions, producing a 1536-dimensional combined representation.

These fused embeddings are then processed through a Multi-Head Attention Transformer (Vaswani et al., 2017) that enables deep interaction between modalities at multiple representation levels. Our models are evaluated across all four subtasks of the CASE 2025 competition to assess their robustness, transferability, and domain-independence. The experimental results validate the effectiveness of our early-fusion attention-based approach in capturing nuanced multimodal intent, outperforming or matching task-specific baselines while maintaining general applicability.

3.1 Architecture 1 : Transformer-based Fusion

This section describes the architecture and training procedure of our unified multimodal classification framework.This approach leverages pretrained encoders for independent modality processing and combines their outputs using a Transformer-based fusion module.

3.1.1 Text Encoding with XLM-RoBERTa

We process the OCR-extracted text from memes using the multilingual transformer XLM-RoBERTa (Conneau et al., 2019). Given an input text sequence $\mathbf{x}_{\text{text}}$, we obtain its contextual representation via the final [CLS] token embedding:

$$\mathbf{h}_{\text{xlmr}} = \text{XLM-R}(\mathbf{x}_{\text{text}})_{[\text{CLS}]} \in \mathbb{R}^{768}$$

This text encoder is fine-tuned using binary cross-entropy loss:

$$\mathcal{L}_{\text{text}} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)$$

where $\hat{y}_i = \sigma(\mathbf{W}_t \cdot \mathbf{h}_{\text{xlmr}} + b_t)$ is the predicted probability of the positive class.

Table 1: CASE 2025 Shared Task Subtasks and Labels

Subtask	Objective	Label Names	Encoded Labels
A	Detect hate speech in the meme.	No Hate, Hate	0, 1
B	Identify the target of hateful memes.	Undirected, Individual, Community, Organization	0, 1, 2, 3
C	Determine stance toward movements.	Neutral, Support, Oppose	0, 1, 2
D	Detect humor/satire/sarcasm.	No Humor, Humor	0, 1

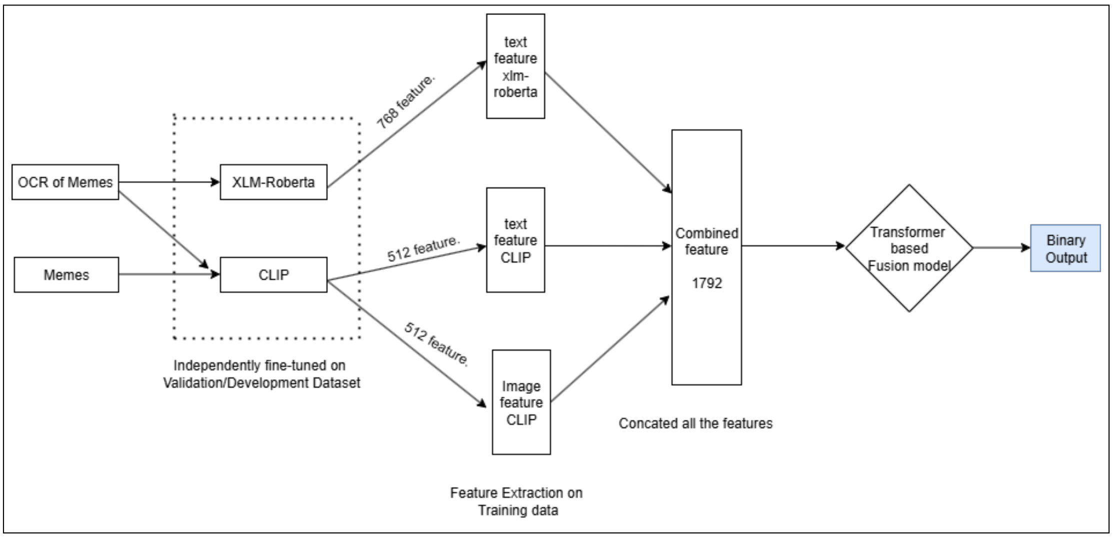


Figure 2: Block diagram of the multimodal architecture for HM, MAMI, and MultiOFF datasets.

3.1.2 Visual and Textual Embedding with CLIP

We employ the Contrastive Language–Image Pretraining (CLIP) model (Radford et al., 2021) to obtain aligned embeddings for both the meme image $\mathbf{x}_{\text{img}}$ and its OCR-extracted text $\mathbf{x}_{\text{ocr}}$. CLIP provides a joint representation:

$$\mathbf{h}_{\text{clip}} = \left[\mathbf{h}_{\text{img}}; \mathbf{h}_{\text{ocr}}\right] \in \mathbb{R}^{1024}, \quad \mathbf{h}_{\text{img}}, \mathbf{h}_{\text{ocr}} \in \mathbb{R}^{512}$$

Both image and text features are extracted via CLIP's pretrained encoders and optionally fine-tuned using:

$$\mathcal{L}_{\text{clip}} = -y \log \sigma(\mathbf{W}_c \cdot \mathbf{h}_{\text{clip}} + b_c)$$

3.1.3 Feature Fusion and Classification

The output representations from XLM-R and CLIP are concatenated into a single vector:

$$\mathbf{z} = [\mathbf{h}_{\text{xlmr}}; \mathbf{h}_{\text{img}}; \mathbf{h}_{\text{ocr}}] \in \mathbb{R}^{1792}$$

This vector is linearly projected to a reduced dimension d for input into a Transformer encoder:

$$\mathbf{z}_{\text{proj}} = \mathbf{W}_p \cdot \mathbf{z} + \mathbf{b}_p, \quad \mathbf{W}_p \in \mathbb{R}^{d \times 1792}$$

The Transformer encoder $\mathcal{T}$ with positional encodings captures inter-modal interactions:

$$\mathbf{z}_{\text{fused}} = \mathcal{T}(\text{PosEnc}(\mathbf{z}_{\text{proj}}))$$

A multilayer perceptron (MLP) followed by a sigmoid activation performs final classification:

$$\hat{y} = \sigma(\text{MLP}(\mathbf{z}_{\text{fused}}))$$

3.1.4 Training Configuration

The fusion classifier is trained using the following setup: Optimizer: Adam, Learning rate: 1×10^{-4}, Loss function: Binary Cross-Entropy, Batch size: 16, and Epochs: 100. The final prediction $\hat{y}$ is thresholded at 0.4:

$$\hat{y} = \begin{cases} 1, & \text{if } \sigma(f(\mathbf{z})) \geq 0.4 \\ 0, & \text{otherwise} \end{cases}$$

3.2 Architecture 2 : Transformer-based Fusion with Bidirectional Cross Attention

3.2.1 Text Encoder: XLM-RoBERTa

The textual content from memes is tokenized and passed into a fine-tuned **XLM-RoBERTa** model. We extract contextual token embeddings $\mathbf{H}_t \in$

$\mathbb{R}^{L\times d}$ and apply mean pooling to obtain the final text embedding:

$$\mathbf{h}_t = \frac{1}{L}\sum_{i=1}^{L}\mathbf{H}_t^{(i)}$$

where L is the sequence length and $d = 768$ is the hidden dimension.

3.2.2 Image Encoder: Vision Transformer (ViT)

The meme image is resized and fed into a pretrained **ViT** model. The image is split into P patches and encoded to obtain $\mathbf{H}_v \in \mathbb{R}^{P\times d}$. Similar to the text stream, we perform mean pooling:

$$\mathbf{h}_v = \frac{1}{P}\sum_{j=1}^{P}\mathbf{H}_v^{(j)}$$

3.2.3 Bidirectional Cross-Modal Attention

We apply **bidirectional multi-head attention** between visual and textual sequences to model fine-grained interactions:

$$\mathbf{A}_{t\leftarrow v} = \text{MHA}(\mathbf{H}_t, \mathbf{H}_v, \mathbf{H}_v) \quad (1)$$

$$\mathbf{A}_{v\leftarrow t} = \text{MHA}(\mathbf{H}_v, \mathbf{H}_t, \mathbf{H}_t) \quad (2)$$

These attention outputs are pooled to form final fused features:

$$\mathbf{z} = [\text{MeanPool}(\mathbf{A}_{t\leftarrow v}); \text{MeanPool}(\mathbf{A}_{v\leftarrow t})] \in \mathbb{R}^{2d}$$

3.2.4 Multi-task Classification Heads

The fused vector $\mathbf{z}$ is passed through four independent multi-layer perceptrons (MLPs), one for each subtask:

$$\hat{\mathbf{y}}^{(s)} = \text{Softmax}(f^{(s)}(\mathbf{z})), \quad s \in \{A, B, C, D\}$$

Each head uses a cross-entropy loss:

$$\mathcal{L}_{\text{total}} = \sum_s \lambda_s \cdot \mathcal{L}_{\text{CE}}^{(s)}$$

where λ_s are task-specific weights (default 1.0).

3.2.5 Training Configuration

Following hyperparameter were used to train the proposed model: Optimizer: AdamW with weight decay, Learning rate: 3×10^{-5} (with warm-up), Epochs: 30, Loss: Multi-task CrossEntropy, Backbone Freezing: First 5 epochs.

3.3 Architecture 3 : Lightweight Fusion with PCA and Multi-Head Attention

3.3.1 Static Feature Extraction

In this architecture, we use frozen encoders for feature extraction:

- **Text:** BERT [CLS] embeddings $\mathbf{t} \in \mathbb{R}^{768}$
- **Image:** ViT global embeddings $\mathbf{v} \in \mathbb{R}^{768}$

3.3.2 Dimensionality Reduction

We apply PCA separately:

$$\mathbf{t}' = \text{PCA}(\mathbf{t}) \in \mathbb{R}^{128}, \quad \mathbf{v}' = \text{PCA}(\mathbf{v}) \in \mathbb{R}^{128}$$

3.3.3 Multi-Head Attention Fusion

The reduced embeddings are passed through dense layers and fused using multi-head attention:

$$\mathbf{q} = \mathbf{W}_q \cdot \mathbf{v}' \quad \mathbf{k}, \mathbf{v}_{att} = \mathbf{W}_k \cdot \mathbf{t}' \quad (3)$$

$$\mathbf{z}_{\text{fused}} = \text{MHA}(\mathbf{q}, \mathbf{k}, \mathbf{v}_{att}) \quad (4)$$

3.3.4 Multi-task Output and Loss

The attention output is flattened and passed into shared dense layers, followed by task-specific classification heads:

$$\hat{\mathbf{y}}^{(s)} = \text{Softmax}(g^{(s)}(\mathbf{z}_{\text{fused}})), \quad \forall s \in \{A, B, C, D\}$$

We minimize total cross-entropy loss across all subtasks:

$$\mathcal{L}_{\text{total}} = \sum_s \mathcal{L}_{\text{CE}}^{(s)}$$

3.3.5 Training Configuration

Following hyperparameter were used to train the proposed model: Optimizer: Adam, Learning Rate: 1×10^{-4}, Epochs: 50, Loss: Cross-entropy per subtask.

4 Results

Table 2 presents the performance of our multimodal model (XLM-R/BERT + ViT/CLIP) across the four subtasks defined in the CASE 2025 shared task. The model achieves reasonably competitive performance in binary tasks such as Hate Detection (Subtask A) and Humor Detection (Subtask D), obtaining an F1 score of 0.6602 and 0.6564, respectively. These results indicate that the model is

Table 2: Performance of the Best Performed Models for Different Sub-tasks

Tasks	**Model**	**Recall**	**Precision**	**F1 Score**	**Accuracy**
Subtask-A	XLM-R + ViT + Attention	0.6614	0.6654	0.6602	0.6627
Subtask-B	BERT + ViT + Attention	0.3158	0.2739	0.4096	0.4217
Subtask-C	XLM-R + ViT + Attention	0.4723	0.4905	0.4674	0.4694
Subtask-D	XLM-R + ViT + Attention	0.6554	0.6575	0.6564	0.7002

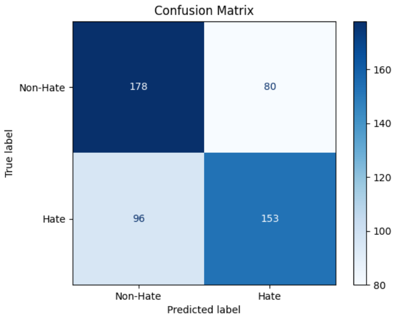


Figure 3: Confusion matrix Subtask A

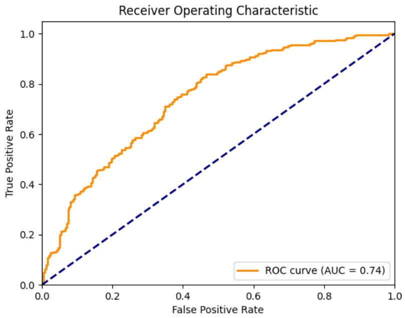


Figure 4: ROC curve multiclass subtask A

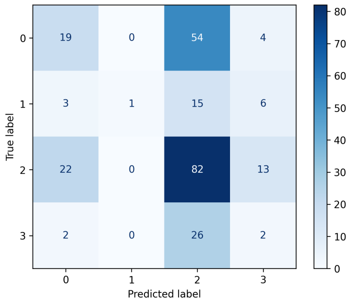


Figure 5: Confusion matrix Subtask B

able to capture surface-level multimodal features to some extent.

However, the performance significantly drops in more semantically complex tasks particularly in Target Identification (Subtask B) and Stance Classification (Subtask C). For instance, the F1 score for Subtask B was only 0.4096, far from the top performer (0.6530). This gap indicates challenges in understanding nuanced and context-specific semantic relationships.

A key reason behind the lower performance in these subtasks is likely due to semantic misalignment between the textual and visual streams. The model often misinterprets the context when the image and text convey conflicting or sarcastic messages. Since memes are frequently designed with contradictory visuals and text (e.g., humorous images paired with hateful text or vice versa), the fusion mechanism occasionally diverges in learning, either overemphasizing the visual cue or misjudging the intended sentiment of the text.

The confusion metric and ROC for for the testing labels for xlm-roberta + VIT+ Attention model can be seen in Figures 3 and 4, respectively. The confusion metric and ROC for for the testing labels for BERT + VIT+ Attention model can be seen in Figures 5 and 6, respectively. The confusion metric and ROC for for the testing labels for clm-roberta + CLIP+ Attention based fusion model can be seen in Figures 7 and 8, respectively. Similarly, confusion matrix and ROC curve for the subtask-D can be seen in Figures 9 and 10, respectively.

These findings echo observations from prior works on multimodal sarcasm and hate detection (Fersini et al., 2022),which highlight that surface-level fusion techniques are insufficient when the modalities encode different or even conflicting semantic signals. Thus, future iterations of the model can benefit from deeper semantic alignment modules or attention-based conflict resolution strategies to better handle such intricacies.

5 Conclusion

The proposed model is validated for four subtasks from the CASE 2025 shared task. The system showed promising performance on binary classification tasks like Hate Detection (F1 = 0.6602) and Humor Detection (F1 = 0.6564), indicating that the model can capture explicit cues from both modalities effectively. However, for more semantically complex subtasks such as Target Group Identification and Stance Classification, the performance was notably lower (F1 = 0.4096 and 0.4674 respectively). These tasks require a deeper understanding of socio-political context and subtle narrative tones, which our current model struggled to generalize. One key limitation identified was the semantic misalignment between image and text. The model often failed to resolve contradictions in multimodal

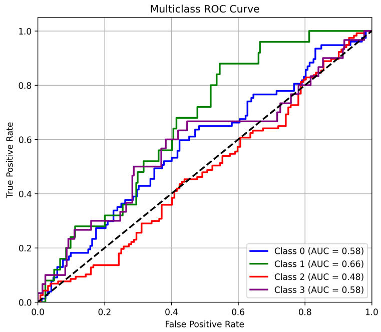


Figure 6: ROC curve for the Subtask B

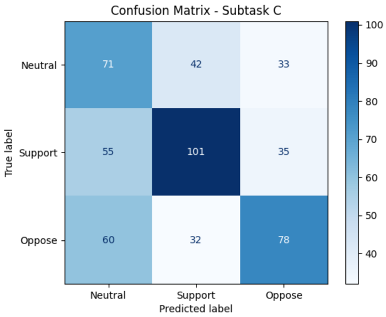


Figure 7: Confusion matrix Subtask C

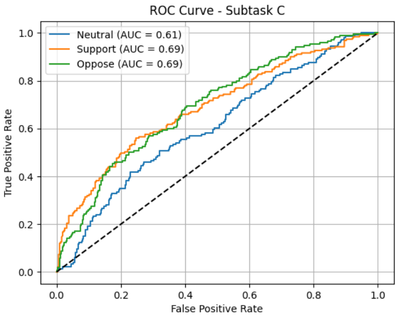


Figure 8: AUC curve multiclass subtask C

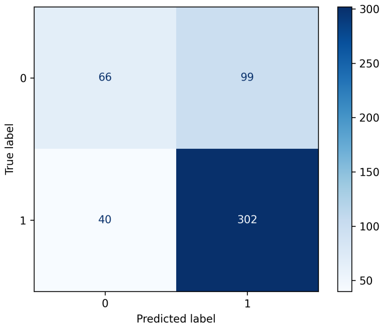


Figure 9: Confusion matrix Subtask D

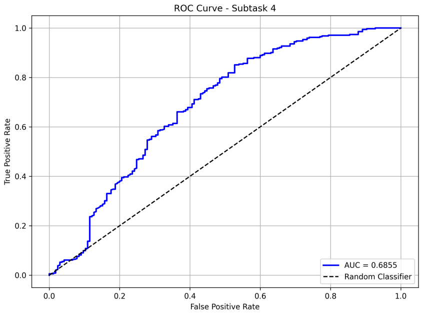


Figure 10: AUC curve multiclass subtask D

memes—where visual irony or sarcasm alters the literal meaning of the text. This led to misinterpretation in scenarios where the intended sentiment was obfuscated through meme-specific humor or design.

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.

Shraddha Chauhan and Abhinav Kumar. 2025. MNLP@DravidianLangTech 2025: transformer-based multimodal framework for misogyny meme detection. In *DravidianLangTech*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Samir El-amrany, Salima Lamsiyah, Matthias R Brust, and Pascal Bouvry. 2025. Guardharmem and harmdetect: a multimodal dataset and benchmlark model for fine-grained harmful meme classification. *Social Network Analysis and Mining*, 15(1):63.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Eftekhar Hossain, Omar Sharif, Moshiul Hoque, and Sarah M. Preum. 2024. Align before attend: Aligning visual and textual features for multimodal hateful content detection. *arXiv preprint arXiv:2402.09738*.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE* *conference on computer vision and pattern recognition*, pages 4700–4708.

Jinfa Huang, Jinsheng Pan, Zhongwei Wan, Hanjia Lyu, and Jiebo Luo. 2024. Evolver: Chain-of-evolution prompting to boost large multimodal models for hateful meme detection. *arXiv preprint arXiv:2407.21004*.

Ali Hürriyetoğlu, Surendrabikram Thapa, Hristo Tanev, and Surabhi Adhikari. 2025. Findings and insights from the 8th workshop on challenges and applications of automated extraction of socio-political events from text. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2025)*.

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.

Abhinav Kumar, Pradeep Kumar Roy, and Sunil Saumya. 2021. An ensemble approach for hate and offensive language identification in english and indo-aryan languages. In *FIRE (Working Notes)*, pages 439–445.

Rohit Kumar and Sathappan Nandakumar. 2022. Hate-clipper: Multimodal hateful meme classification using vision-language pretraining. *arXiv preprint arXiv:2210.05916*.

Gitanjali Kumari, Kirtan Jain, and Asif Ekbal. 2024. M3hop-cot: Misogynous meme identification with multimodal multi-hop chain-of-thought. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22105–22138.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM Web Conference 2024*, pages 2359–2370.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Jingbiao Mei, Jinghong Chen, Guangyu Yang, Weizhe Lin, and Bill Byrne. 2025. Robust Adaptation of Large Multimodal Models for Retrieval Augmented Hateful Meme Detection. *arXiv e-prints*, page arXiv:2502.13061.

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, Pam Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning (ICML)*.

Pradeep Kumar Roy and Abhinav Kumar. 2025. Ensuring safety in digital spaces: Detecting code-mixed hate speech in social media posts. *Data & Knowledge Engineering*, 156:102409.

Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. 2024. Memeclip: Leveraging clip representations for multimodal meme classification. pages 17320–17332.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas Pykl, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. Semeval-2020 task 8: Memotion analysis-the visuo-lingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773.

Manswini Swain, Manish Biswal, Priya Raj, Abhinav Kumar, and Debahuti Mishra. 2022. Hate and offensive language identification from social media: a machine learning approach. In *Electronic Systems and Intelligent Computing: Proceedings of ESIC 2021*, pages 335–342. Springer.

Surendrabikram Thapa, Farhan Ahmad Jafri, Ali Hürriyetoğlu, Francielle Vargas, Roy Ka Wei Lee, and Usman Naseem. 2023. Multimodal hate speech event detection-shared task 4. In *CASE 2023-Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text, associated with 14th International Conference on Recent Advances in Natural Language Processing, RANLP 2023*, pages 151–159. Association for Computational Linguistics.

Surendrabikram Thapa, Kritesh Rauniyar, Farhan Ahmad Jafri, Hariram Veeramani, Raghav Jain, Sandesh Jain, Francielle Vargas, Ali Hürriyetoğlu, and Usman Naseem. 2024. Extended multimodal hate speech event detection during russia-ukraine crisis-shared task at case 2024. In *7th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text, CASE 2024*, pages 221–228. Association for Computational Linguistics.

Surendrabikram Thapa, Siddhant Bikram Shah, Kritesh Rauniyar, Shuvam Shiwakoti, Surabhi Adhikari, Hariram Veeramani, Kristina T. Johnson, Ali Hürriyetoğlu, Hristo Tanev, and Usman Naseem. 2025a. Multimodal hate, humor, and stance event detection in marginalized sociopolitical movements. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Sociopolitical Events from Text (CASE 2025)*.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the origins of memes by means of fringe web communities. In *Proceedings of the internet measurement conference 2018*, pages 188–202.