PhantomTroupe@CASE 2025: Multimodal Hate Speech Detection in Text-Embedded Memes using Instruction-Tuned LLMs

Farhan Amin, Muhammad Abu Horaira, Md. Tanvir Ahammed Shawon,
Md Ayon Mia, Muhammad Ibrahim Khan
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u2004068, u2004029, u1904077,u1804128}@student.cuet.ac.bd, muhammad_ikhan@cuet.ac.bd

Abstract

Memes and other text-embedded images are powerful tools for expressing opinions and identities, especially within marginalized socio-political movements. Detecting hate speech in this type of multimodal content is challenging because of the subtle ways text and visuals interact. In this paper, we describe our approach for Subtask A of the Shared Task on Multimodal Hate Detection in Marginalized Movement@CASE 2025, which focuses on classifying memes as either Hate or No Hate. We tested both unimodal and multimodal setups, using models like DistilBERT, HateBERT, Vision Transformer, and Swin Transformer. Our best system is the large multimodal model Qwen2.5-VL-7B-Instruct-bnb-4bit, fine-tuned with 4-bit quantization and instruction prompts. While we also tried late fusion with multiple transformers, Qwen performed better at capturing text-image interactions in memes. This LLM-based approach reached the highest F1-score of 0.8086 on the test set, ranking our team 5th overall in the task. These results show the value of late fusion and instruction-tuned LLMs for tackling complex hate speech in socio-political memes.

1 Introduction

Social media has become a fast-paced platform where content spreads instantly, with memes playing a big role in communication. But they are also used to spread harmful messages, including hate speech targeting marginalized groups. This kind of content can make online spaces unsafe.Since it is impossible to manually keep up with everything being shared, an automated system has become essential for managing such content. This paper addresses Subtask A of the Shared Task on Multimodal Hate Detection in Marginalized Movement@CASE2025, focusing on binary classification ('No Hate' vs. 'Hate') across a dataset of 4,675 text-embedded images. We draw inspiration from (Parihar et al., 2021), which explores natural language processing for identifying harmful content, shaping our approach to this challenge. To address this challenge, we developed a practical approach by fine-tuning transformer models such as DistilBERT and HateBERT to capture textual nuances, and enhancing it with late fusion to integrate visual data, where Qwen2.5-VL-7B-Instruct-bnb-4bit showed strong capability in interpreting the socio-political nuances of memes. This aligns with (Thapa et al., 2025b), which highlights LLM potential in social science, and builds on (Thapa et al., 2023) and (Chhabra and Vishwakarma, 2024) for multimodal insights. Through this work, we hope to contribute towards more scalable and fair content moderation solutions. Our main contributions are as follows:

- A systematic comparison of unimodal, late-fusion multimodal, and LLM-based architectures for meme hate speech detection.
- An efficient fine-tuning strategy that combines LoRA with 4-bit quantization to adapt a large multimodal LLM under resource constraints.
- Empirical analysis of model predictions, illustrated with representative examples drawn from different regions of the confusion matrix.

2 Related Works

Previous research on multimodal hate speech detection has explored many creative ways to tackle the challenges of online conversations, especially in complex social and political contexts. Early work like (Parihar et al., 2021) used natural language processing to spot hate speech by looking at language patterns that show harmful intent. Later studies, such as (Kashif et al., 2023), used ensemble learning to combine features from different data types for better results. Similarly, (Sahin et al., 2023) improved text analysis by adding syntactic and entity-

Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025, pages 133–138, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-016

level information with transformer models. In another approach, (Aziz et al., 2023) proposed a hierarchical fusion method with separate transformer encoders, and (Chhabra and Vishwakarma, 2024) developed a scalable multilevel attention framework that has influenced our work. While these studies built a strong base for cross-modal hate detection, many still require heavy computation and can be hard to interpret, especially when handling satire or cultural references in memes.

Shared tasks have also helped shape this field by providing benchmarks and valuable datasets. The Shared Task on Multimodal Hate Detection in Marginalized Movement@CASE2025, described in (Thapa et al., 2025a), focuses on binary hate classification, with insights from (Hürriyetoğlu et al., 2025) showing how it has grown. The CASE 2024 shared task (Thapa et al., 2024) featured participants demonstrating the utilization of transformer models like BERT, RoBERTa, and XLNet, as well as effective approaches such as vision transformers and CLIP, which contributed to the outstanding outcomes in hate event detection. The CASE 2023 shared task (Thapa et al., 2023) laid the groundwork for multimodal hate speech detection by combining textual and visual features in text-embedded images

Our dataset comes mainly from (Shah et al., 2024) and its CLIP-based representations, supported by (Bhandari et al., 2023)'s CrisisHateMM work, which highlights the value of careful data curation. Finally, (Thapa et al., 2025b) discusses how large language models can help in social science research, encouraging us to tackle ongoing challenges like telling satire apart from hate in fast-changing socio-political memes.

3 Task and Dataset Description

We have utilized the dataset provided for Subtask A of the Shared Task on Multimodal Hate Detection in Marginalized Movement@CASE2025, as outlined in (Thapa et al., 2025a), focusing on detecting hate speech in text-embedded images. The dataset is divided into training, validation, and test sets with 3,662, 506, and 507 samples, respectively, primarily comprising memes and similar online images. Each image is labeled with a binary tag: 'Hate' or 'No Hate', as detailed in Table 1. This dataset, curated for the 2025 task, serves as our primary resource, with its development informed by (Shah et al., 2024) for CLIP-based representations and supplemented by (Bhandari et al., 2023) for CrisisHateMM analysis, which also shapes the annotation schema.

Table 1: Distribution of images for Subtask A meme hate speech detection.

Dataset	No Hate	Hate	Total
Train	1930	1732	3662
Val	258	248	506
Test	258	249	507

The relatively small dataset size presents challenges for training transformer models, as it increases the risk of overfitting, which motivated our use of data augmentation.

4 Methodology

4.1 Preprocessing

As this is a multimodal task, we have preprocessed both text and image. For the text, we have removed URLs, HTML tags, emojis, and extra whitespace to reduce noise, and converted all text to lowercase for consistency. On the image side, all samples were converted to RGB, resized to 224×224 pixels, and normalized using ImageNet mean and standard deviation to match the input requirements of pretrained models. In total, we have found that 3,662 out of 4,050 training samples had images that matched the text, and we have discarded the rest. All 506 validation samples and all 507 test samples had no missing images.

4.2 Augmentation

To improve model generalization and reduce overfitting, we have applied data augmentation techniques during training. Each image was randomly flipped horizontally and cropped with padding to introduce variation while preserving semantic content. These augmentations were applied only to the training set, while the validation and test sets were left unchanged to ensure consistent evaluation.

4.3 Transformer-based Approach

4.3.1 Unimodal Approach

For the unimodal text classification task, we fine-tuned two transformer-based models: DistilBERT-base-uncased and GroNLP/HateBERT. We selected these models for their pretrained knowledge of general language and hate speech domains. Text sequences were tokenized with a maximum length of 128 tokens. We included a dropout rate of 0.2 in the hidden and attention layers to help prevent overfitting. We trained the models using the Adam

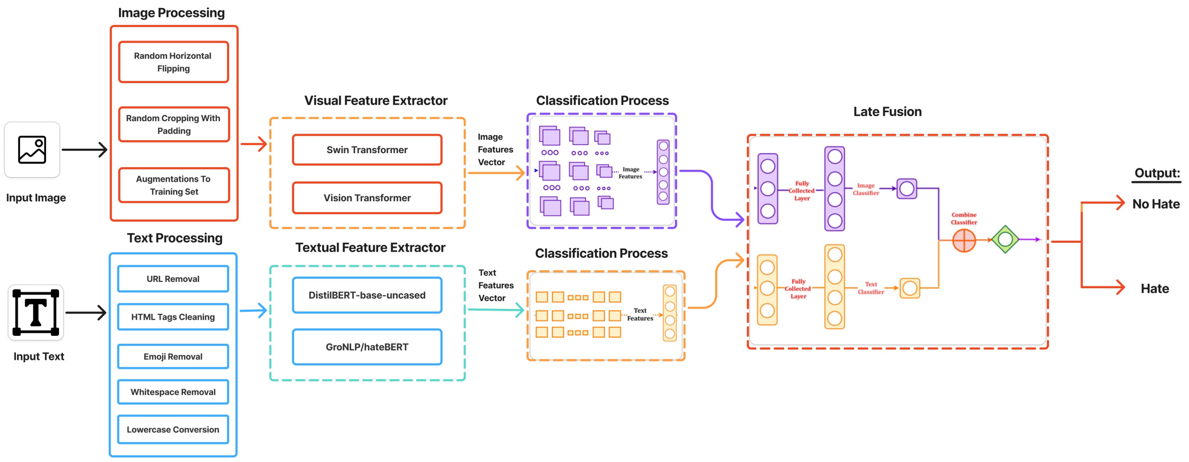


Figure 1: Transformer-based multimodal late fusion architecture for meme hate speech detection

optimizer with a learning rate of 2×10^{-5}, a weight decay of 0.01, and ran the training for 10 epochs on the provided training set. We tokenized the validation and test sets in the same way for evaluation.

For the image-only models, we experimented with Vision Transformer (ViT-base-patch16-224) and Swin Transformer (Swin-T-patch4-window7-224), both of which were first trained on ImageNet. We resized the input images to 224×224, normalized them with standard ImageNet statistics, and converted them to tensors. Each model extracted a 768-dimensional feature vector from the images. We added a dropout layer and a classification head to predict binary labels for Hate and No Hate. Both models were trained for 10 epochs using the AdamW optimizer with a learning rate of 2×10^{-5} and a batch size of 16.

4.3.2 Multimodal Approach

Building on the unimodal baselines, we developed two multimodal architectures that combine both text and images: Swin Transformer with DistilBERT, and HateBERT with Vision Transformer (ViT). In both configurations, 768-dimensional embeddings were extracted separately from the image and text inputs. We combined the embeddings using a late-fusion approach by simply concatenating them. This was followed by a dropout layer to reduce overfitting and a final linear layer for classification. Among the two, the Swin + DistilBERT combination consistently achieved the best performance on the test set.

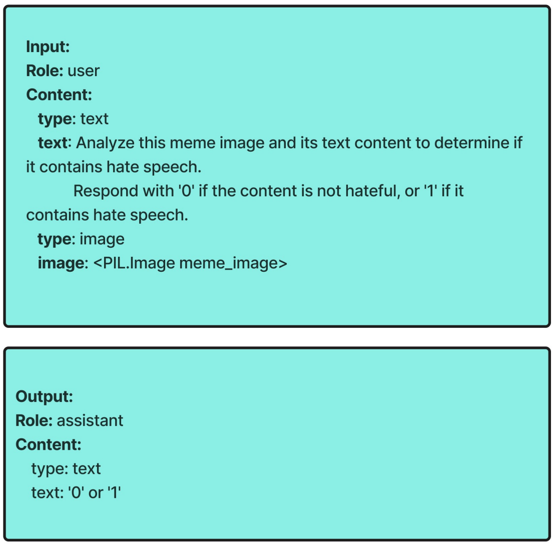
Late fusion performed better as it enabled the model to process images and text independently before combining their representations, allowing each modality to contribute its strengths. This separation allowed each type of data to focus on its strengths, like visual features from the image and contextual meaning from the text. Combining them later helps the model pick up on subtle clues that come from both. This is really important in hate speech detection, where sometimes the meaning hides in the image, sometimes in the text, and often in both together.

4.4 LLM-Based Approach

We employed a multimodal large language model, `Qwen2.5-VL-7B-Instruct-bnb-4bit`, fine-tuned using the Unsloth framework with 4-bit quantization to improve training efficiency. The goal was to detect hate speech in memes by analyzing both their visual and textual content together. Each training instance was structured as a chat-style conversation, where the user provides an instruction along with a meme image, and the assistant outputs either `0` or `1`, indicating the absence or presence of hate speech, respectively.

We fine-tuned the model over 7 epochs with a batch size of 32. We used LoRA-based fine-tuning (Low-Rank Adaptation) with a rank of 128 applied to both vision and language components. During inference, we applied a zero-shot prompting strategy by employing the same instruction without any meme-specific customization and constrained the model to generate a single classification token.

Our approach achieved a test F1-score of 0.8086, demonstrating efficient performance without relying on handcrafted prompts. This highlights how effective and scalable instruction-tuned multimodal large language models are for detecting hate speech.



Input:
Role: user
Content:
type: text
text: Analyze this meme image and its text content to determine if it contains hate speech.
Respond with '0' if the content is not hateful, or '1' if it contains hate speech.
type: image
image: <PIL.Image meme_image>

Output:
Role: assistant
Content:
type: text
text: '0' or '1'

Figure 2: Prompt provided to Qwen2.5-VL-7B-Instruct-bnb-4bit for meme hate speech detection.

We selected Qwen2.5-VL because it is open-source with efficient fine-tuning support, provides strong multimodal reasoning comparable to larger closed-source models such as GPT-4V, and supports quantized training, where we adopted 4-bit due to hardware limitations with 8-bit.

5 Results and Analysis

5.1 Comparative Analysis

Between the two unimodal text classifiers, HateBERT outperformed DistilBERT-base-uncased, achieving a higher F1-score of 0.7810 compared to 0.7424. This indicates that HateBERT is more effective for hate speech detection on meme texts, likely because it is pretrained specifically on hate speech data. For the unimodal image models, Swin Transformer outperformed ViT, achieving 0.6668 compared to 0.6166, indicating stronger visual feature extraction.

In the multimodal setups, we adopted a late fusion strategy to combine textual and visual representations. Using this approach, Swin Transformer + DistilBERT achieved an F1-score of 0.7790, slightly outperforming the ViT + HateBERT model which scored 0.7576. These results highlight how late fusion enables each modality to contribute its strengths independently before combining them for final prediction, leading to better performance than unimodal baselines.

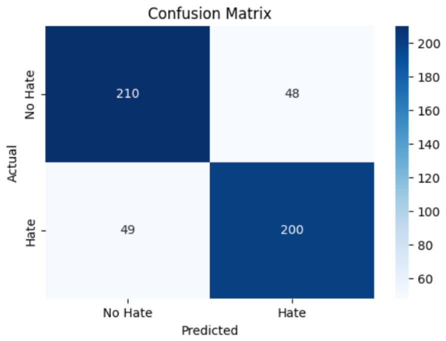


Figure 3: Confusion matrix showing the proposed model's binary classification performance for meme hate speech detection.

Finally, the best overall performance was obtained by fine-tuning Qwen2.5-VL-7B-Instruct-bnb-4bit using the Unsloth framework. This model achieved an F1-score, precision, and recall of 0.8086. These results highlight the strong potential of large-scale instruction-tuned multimodal models in capturing subtle and cross-modal patterns in hateful memes, outperforming both traditional and multimodal baselines. The results are detailed in Table 2.

5.2 Error Analysis

To better understand the limitations of our best model, we examined its confusion matrix. The fine-tuned Qwen2.5-VL-7B-Instruct-bnb-4bit model correctly predicted most instances in both classes, but some misclassifications remain. It falsely labeled 48 non-hateful memes as hateful and misclassified 49 actual hateful ones.

These errors suggest that while the model performs well overall, it occasionally struggles with subtle or ambiguous cases where hateful intent is not explicit.

5.3 Quantitative Analysis

The confusion matrix shows a fairly balanced distribution of errors, with 48 false positives and 49 false negatives. This indicates that the model is not heavily biased toward one class. However, the nearly equal misclassifications suggest that the model may still be relying on surface-level features, such as specific keywords or visual patterns, rather than understanding the deeper context. Exploring techniques like attention visualization or feature attribution could help reveal what the model is focusing on and guide improvements in handling more nuanced or borderline cases.

5.4 Qualitative Analysis

To further investigate the model's decision patterns, we sampled representative examples from each con-

Table 2: Performance comparison of different models for meme hate speech detection.

Classifier	P	R	F1	Accuracy
Unimodal (Text)				
HateBERT	0.7810	0.7811	0.7810	0.7811
DistilBERT-base-uncased	0.7424	0.7424	0.7424	0.7424
Unimodal (Image)				
Vision Transformer (ViT)	0.6166	0.6116	0.6085	0.6134
Swin Transformer	0.6668	0.6661	0.6660	0.6667
Multimodal				
Swin Transformer + DistilBERT (Late Fusion)	0.7790	0.7768	0.7792	0.7792
ViT + HateBERT (Late Fusion)	0.7576	0.7576	0.7579	0.7579
LLMs				
Qwen2.5-VL-7B-Instruct-bnb-4bit	**0.8086**	**0.8086**	**0.8086**	**0.8087**

fusion matrix category:

Table 3: Example predictions illustrating each category of the confusion matrix.

Category	Example (Image)
True Positive (TP)	1040.png
True Negative (TN)	1155.png
False Positive (FP)	1011.png
False Negative (FN)	1002.png

As shown in Table 3, the model performs well when hateful intent is **clear and explicit**. For instance, it correctly labels a public gathering with signs promoting fairness and unity as **No Hate**, and it also identifies **explicit hostility** in text-based images, such as content expressing negativity toward a music genre.

The model struggles more with memes that are **ambiguous or context-dependent**. A false positive example, a meme satirizing corporate behavior during awareness campaigns, was incorrectly flagged as **Hate**, showing difficulty in separating satire from genuine hostility. Similarly, a false negative case, a humorous meme about dating, was misclassified as **No Hate**, reflecting the challenge of detecting humor that may conceal harmful undertones.

Overall, these patterns highlight the need for stronger cross-modal reasoning and better interpretability to handle subtle and context-driven cases.

6 Conclusion

In this study, we tackled the challenge of detecting hate speech in text-embedded images as part of the Shared Task on Multimodal Hate Detection in Marginalized Movement@CASE2025, focusing on Subtask A. We used a fine-tuned Qwen2.5-VL-7B-Instruct-bnb-4bit model combined with a late-fusion strategy to merge textual and visual features. This approach achieved a solid F1-score of 0.8086 on the test set, highlighting the model's ability to capture subtle interactions between modalities for spotting hate speech in complex memes and socio-political contexts. When compared with unimodal and other multimodal baselines, our method showed clear improvements, especially when humor and harmful messages are mixed together. Overall, our findings provide a useful approach for handling multimodal content, particularly where it relates to marginalized groups.

7 Limitations

We selected Qwen2.5-VL with 4-bit quantization because it is openly available, resource-efficient, and feasible within our computational constraints. However, stronger models (e.g., BLIP-2, LLaVA, GPT-4V) and higher-precision training could potentially yield better results. A single fixed prompt was used for simplicity, though alternative prompting strategies or retrieval-based methods may improve robustness. We adopted late fusion for efficiency, but more advanced cross-modal fusion techniques could capture interactions more effectively. Finally, the dataset (4,675 samples) is relatively small, which may limit generalization and reduce coverage of subtle or context-dependent hate speech, highlighting the need for larger datasets in future work.

8 Ethics Statement

We have been committed to ethical practices in developing a system to detect hate speech in images related to marginalized movements. We understand the risks of mislabeling content and worked to balance false positives and negatives, achieving a strong F1-score. Using a public dataset without extra annotation, we respected privacy and data guidelines. Our goal is to promote safer online spaces by reducing harmful content, while recognizing that human oversight is needed to handle context and avoid bias.

References

Abdul Aziz, Md. Akram Hossain, and Abu Nowshed Chy. 2023. Csecu-dsg@multimodal hate speech event detection 2023: Transformer-based multimodal hierarchical fusion model for multimodal hate speech detection. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2023)*, pages 101–107, Varna, Bulgaria. INCOMA Ltd.

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.

Anusha Chhabra and Dinesh Kumar Vishwakarma. 2024. Mhs-stma: Multimodal hate speech detection via scalable transformer-based multilevel attention framework. *arXiv preprint arXiv:2409.05136*.

Ali Hürriyetoğlu, Surendrabikram Thapa, and Hristo Tanev. 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)*.

Mohammad Kashif, Mohammad Zohair, and Saquib Ali. 2023. Lexical squad@multimodal hate speech event detection 2023: Multimodal hate speech detection using fused ensemble approach. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2023)*, pages 84–91, Varna, Bulgaria. INCOMA Ltd.

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.

Umitcan Sahin, Izzet Emre Kucukkaya, Oguzhan Ozcelik, and Cagri Toraman. 2023. Arc-nlp at multimodal hate speech event detection 2023: Multimodal methods boosted by ensemble learning, syntactical and entity features. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2023)*, pages 71–78, Varna, Bulgaria. INCOMA Ltd.

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

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 Socio-political 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.