Team MemeMasters@CASE 2025: Adapting Vision-Language Models for Understanding Hate Speech in Multimodal Content

Shruti Gurung
Patan College for Professional Studies
Kathmandu/Nepal
gurungshrutee44@gmail.com

Shubham Shakya
DoCSE, Kathmandu University
Dhulikhel, Nepal
ss46041720@student.ku.edu.np

Abstract

Social media memes have become a powerful form of digital communication, combining images and text to convey humor, social commentary, and sometimes harmful content. This paper presents a multimodal approach using a fine-tuned CLIP model to analyze text-embedded images in the CASE 2025 Shared Task. We address four subtasks: Hate Speech Detection, Target Classification, Stance Detection, and Humor Detection. Our method effectively captures visual and textual signals, achieving strong performance with precision of 80% for the detection of hate speech and 76% for the detection of humor, while stance and target classification achieved a precision of 60% and 54%, respectively. Detailed evaluations with classification reports and confusion matrices highlight the ability of the model to handle complex multimodal signals in social media content, demonstrating the potential of vision-language models for computational social science applications.

Keywords: Social Media, Memes, Multimodal Analysis, Hate Speech, CLIP

1 Introduction

The explosive rise of social media has transformed memes into powerful tools for both expression and controversy. Memes, text-embedded images that fuse humor, sarcasm, and social commentary, attract millions of users and play an important role in digital culture Arya et al. (2024). They often reflect public sentiment, amplify social trends, and spark dialogue but their layered meanings can also mask harmful intent, making them difficult for researchers to analyze accurately. Studies have shown that even as memes entertain, their content is laden with nuanced signals, necessitating fresh research approaches that integrate visual and linguistic analyses Arya et al. (2024). Recent bibliometric analysis highlights an increasing research interest in the study of memes, underlining their cultural significance and the need for systematic investigation Kamath and Alur (2024). In addition, research on generational humor emphasizes that memes do more than amuse. They also shape identity and social behavior, thus offering valuable insights into emerging cultural dynamics Aronson and Jaffal (2021). This complexity and cultural impact underscore the urgent need for more comprehensive studies that can unravel the multifaceted messages embedded in memes.

To address these complexities, the CASE workshop series introduced shared tasks focused on multimodal analysis of socio-political discourse. Our team participated in the CASE 2025 shared task (Thapa et al., 2025a), which included four subtasks: Hate Speech Detection (A), Target Classification (B), Topical Stance Classification (C), and Intended Humor Detection (D). These build on previous editions, including CASE 2023 (Thapa et al., 2023) and CASE 2024 (Thapa et al., 2024), emphasizing the importance of understanding multimodal online content (Hürriyetoğlu et al., 2025). Each subtask addresses different challenges in the interpretation of complex messages, requiring models to combine textual and visual information for better detection and analysis.

To tackle these challenges, we built on the strengths of modern vision-language models. Specifically, we fine-tuned the openai/clip-vit-large-patch14 model (Radford et al., 2021) to suit each subtask better. This helped the model pick up on subtle signals like sarcasm, implied hostility, and humor which are things that can easily be missed when looking at just text or images alone. By adapting a general-purpose model to these specific tasks, we created a flexible approach for understanding the complex and layered messages found in multimodal online content.

Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025, pages 146–151, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-018

2 Dataset and Task

The experiments described in this paper utilized the dataset provided as part of the CASE 2025 Shared Task on Multimodal Understanding of Online Discourse. This dataset specifically focuses on text-embedded images, such as memes, related to marginalized movements, requiring a nuanced multimodal understanding of the expressions conveyed. The complexity arises from the potential for humor and harm to be intertwined, challenging traditional content moderation approaches. The dataset was created using resources from the Memeclip study (Shah et al., 2024) and earlier multimodal hate speech datasets such as CrisisHateMM (Bhandari et al., 2023), which also contributed to the annotation approach used.

Table 1: Dataset Overview for CASE 2025 Shared Task Subtasks

Subtask	Label	Count	%
ST-A	Non-Hate (0)	2065	51.0
	Hate (1)	1985	49.0
ST-B	Undirected (0)	617	31.1
	Individual (1)	199	10.0
	Community (2)	931	46.9
	Organization (3)	238	12.0
ST-C	Neutral (0)	1166	28.8
	Support (1)	1527	37.7
	Oppose (2)	1357	33.5
ST-D	No Humor (0)	1313	32.4
	Humor (1)	2737	67.5

2.1 Subtask A: Detection of Hate Speech

This subtask aimed to identify the presence of hate speech within text-embedded images. It is framed as a binary classification problem with labels: Non-Hate (0) and Hate (1). The dataset contains a total of 4050 samples, with 2065 (51.0%) labeled as Non-Hate and 1985 (49.0%) labeled as Hate, indicating a relatively balanced distribution (see Table 1).

2.2 Subtask B: Classifying the Targets of Hate Speech

Given an image containing hate speech, the goal of this subtask was to classify the specific target of that hate. This is a multi-class classification problem with four labels: Undirected (0), Individual (1), Community (2), and Organization (3). The dataset includes 1985 samples with notable imbalance: Community targets dominate with 931 samples (46.9%), followed by Undirected (617, 31.1%), Organization (238, 12.0%), and Individual (199, 10.0%) (see Table 1).

2.3 Subtask C: Classification of Topical Stance

This subtask required classifying images based on their stance toward the marginalized movement, with three labels: Neutral (0), Support (1), and Oppose (2). The dataset consists of 4040 samples distributed as follows: Support leads with 1527 samples (37.7%), followed by Oppose at 1357 (33.5%) and Neutral at 1166 (28.8%) (see Table 1), showing a fairly balanced distribution.

2.4 Subtask D: Detection of Intended Humor

The objective here was to identify images conveying humor, sarcasm, or satire related to the marginalized movement. This binary classification task includes labels: No Humor (0) and Humor (1). The dataset is skewed towards humor, with 2737 samples (67.5%) labeled as Humor and 1313 samples (32.4%) labeled as No Humor (see Table 1).

3 Methodology

Our approach across all subtasks was built around the CLIP (Contrastive Language-Image Pre-training) model (Radford et al., 2021), fine-tuned to effectively capture multimodal cues present in text-embedded images. Figure 1 illustrates the overall architecture of our CLIP-based multimodal pipeline, which remained consistent with minor adjustments for each subtask.

3.1 Data Preparation

Each dataset was first parsed and cleaned to ensure valid label mappings according to the task definitions. Text inputs were padded or truncated to CLIP's maximum token length of 77, and images were resized and normalized as per CLIP's preprocessing requirements using the `CLIPProcessor`.

3.2 Dataset and DataLoader

We implemented a custom PyTorch `Dataset` class that dynamically loads paired (image, text) examples and applies the required CLIP-compatible transformations. Batched data was served using a `DataLoader` with shuffling enabled for training and deterministic loading for validation/testing phases.

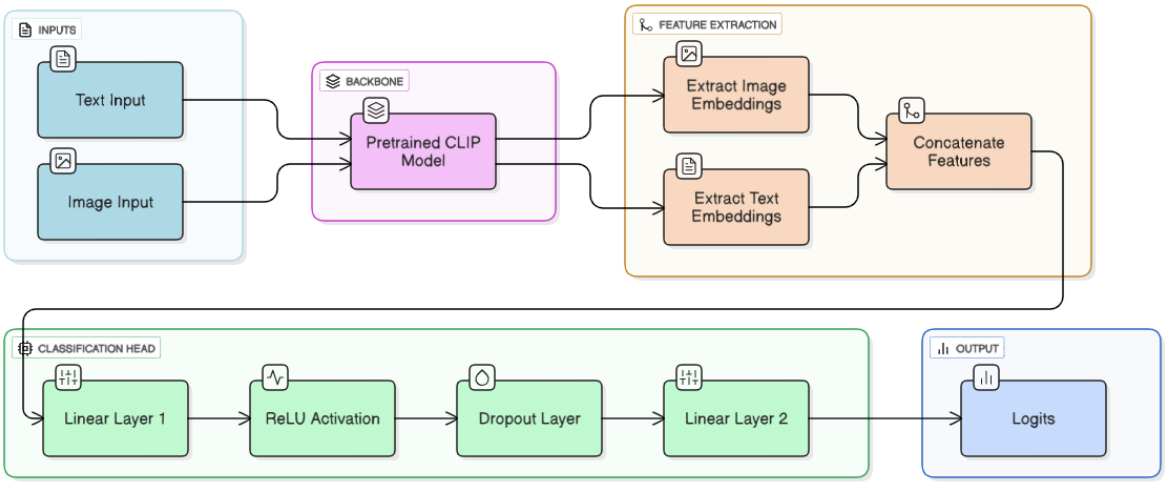

Figure 1: CLIP-Based Multimodal Model Architecture used across all subtasks.

3.3 Model Architecture

The core model utilized the pretrained ViT-L/14 variant of CLIP, where image and text modalities are projected into a shared embedding space. These embeddings were concatenated and passed through a lightweight classification head composed of fully connected layers with ReLU activations and dropout. This head outputted logits over the subtask-specific label set. The architecture is visually depicted in Figure 1.

3.4 Training Procedure

We trained the model using the AdamW optimizer with different learning rates for the backbone and the classification head to facilitate stable fine-tuning. A linear learning rate scheduler with warm-up was used. Training was conducted over 5 epochs with a batch size of 16 using a Tesla T4 GPU. Table 2 summarizes our hyperparameter settings.

3.5 Evaluation and Inference

Performance was monitored using accuracy, precision, recall, and F1-score, with validation conducted at the end of each epoch. The best model checkpoint (based on validation F1-score) was used for generating final predictions on the test set, which were formatted according to the competition submission schema.

3.6 Subtask-Specific Adaptations

While the base setup remained consistent across subtasks, we made targeted modifications where needed. For Subtask B (Target Classification), we applied over-sampling to address class imbalance. Subtask C (Stance Detection) benefited from a deeper 3-layer classifier and a cosine learning rate scheduler instead of linear. For Subtask D (Humor Detection), we used a higher dropout rate and class-weighted loss to handle imbalance. Subtask A (Hate Speech) followed the standard configuration without additional changes.

Table 2: Hyperparameters and Training Configuration

Parameter	Value
Model Backbone	openai/clip-vit-large-patch14
Max Token Length	77
Batch Size	16
Epochs	5
Optimizer	AdamW
Learning Rate (Backbone)	1×10^{-6}
Learning Rate (Classifier Head)	1×10^{-5}
Device	GPU (Tesla T4)
Loss Function	Cross-Entropy

4 Results and Discussion

4.1 Overview

We present evaluation results across CASE 2025 subtasks, with detailed metrics in Table 3 and confusion matrices highlighting common misclassifications. The model performs better on binary tasks like hate speech and humor detection, while multiclass tasks such as stance and target classification remain challenging. These findings reflect known

difficulties in hate speech detection and social media analysis (Parihar et al., 2021).

4.2 Subtask Evaluation

4.2.1 Subtask A: Hate Speech Detection

The model achieved an accuracy of 80% on the binary hate speech detection task, with balanced precision and recall across both classes. As shown in Table 3, the macro-averaged F1-score was 0.80, indicating consistent performance. Class 0 (non-hate) had slightly higher recall (0.82), while class 1 (hate) showed comparable precision (0.81), suggesting cautious detection of hate speech. The confusion matrix in Figure 2 confirms these results, with 212 correctly classified non-hate instances and 194 correctly classified hate instances, alongside 46 false positives and 55 false negatives. Overall, the model performs reliably with minimal bias on this subtask.

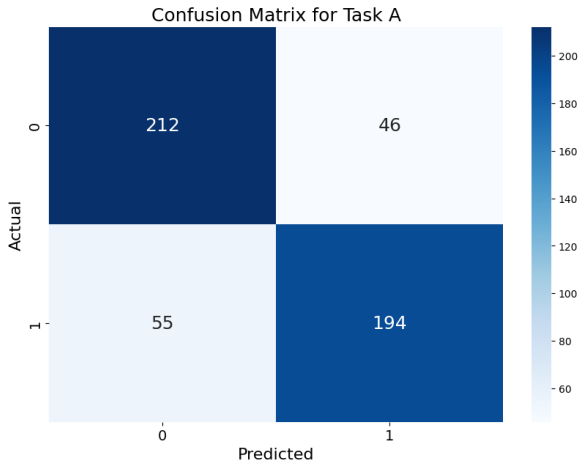

Figure 2: Confusion Matrix for SubTask-A.

4.2.2 Subtask B: Target Classification

For the multi-class classification of hate speech targets, the model achieved an accuracy of 54% as reported in Table 3. Performance varied across classes, with the Community class (2) having the highest recall (0.69) and the Individual class (1) showing the lowest. The confusion matrix in Figure 3 reveals common misclassifications, especially between the Non-Directed (0) and Community (2) classes, indicating some overlap in features. The model handles the imbalanced classes moderately well but struggles with less frequent targets. These results highlight the challenge of fine-grained target detection in hate speech.

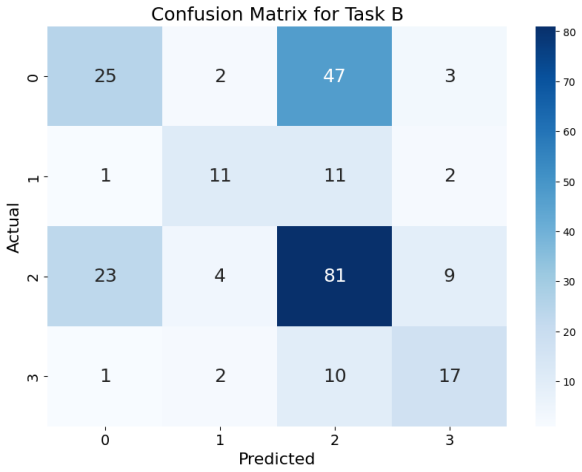

Figure 3: Confusion Matrix for SubTask-B.

4.2.3 Subtask C: Stance Detection

For the multi-class classification of stance, the model achieved an accuracy of 60% as reported in Table 3. Performance varied across classes, with Neutral (0) having the highest recall (0.69), while Support (1) and Oppose (2) were more frequently confused with Neutral. The confusion matrix in Figure 4 shows substantial misclassifications of Support (1) and Oppose (2) as Neutral (0), reflecting the challenge of distinguishing subtle stance differences. Misclassifications between Support (1) and Oppose (2) are also observed, indicating overlap in their features. These results highlight the complexity of stance detection in multimodal online discourse.

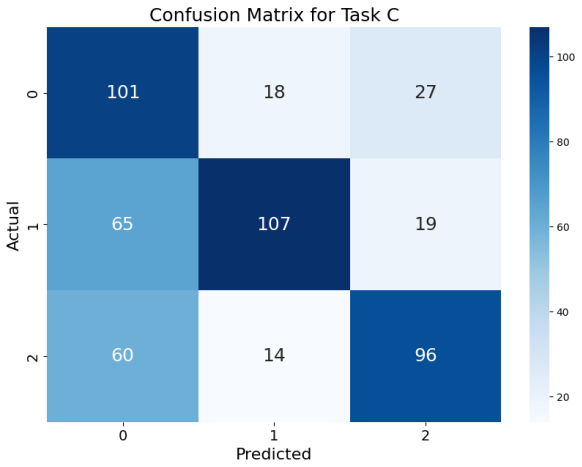

Figure 4: Confusion Matrix for SubTask-C.

4.2.4 Subtask D: Humor Detection

For the binary classification of humor detection, the model achieved an accuracy of 76% as shown

Table 3: Classification Reports for CASE 2025 Subtasks

(a) Subtask A (Hate Speech)

Class (ID)	Precision	Recall	F1-score	Support
Non-Hate (0)	0.79	0.82	0.81	258
Hate (1)	0.81	0.78	0.79	249
Accuracy	0.80			
Macro Avg	0.80	0.80	0.80	507
Weighted Avg	0.80	0.80	0.80	507

(b) Subtask B (Target Classification)

Class (ID)	Precision	Recall	F1-score	Support
Undirected (0)	0.50	0.32	0.39	77
Individual (1)	0.58	0.44	0.50	25
Community (2)	0.54	0.69	0.61	117
Organization (3)	0.55	0.57	0.56	30
Accuracy	0.54			
Macro Avg	0.54	0.51	0.52	249
Weighted Avg	0.53	0.54	0.53	249

(c) Subtask C (Stance Detection)

Class (ID)	Precision	Recall	F1-score	Support
Neutral (0)	0.45	0.69	0.54	146
Support (1)	0.77	0.56	0.65	191
Oppose (2)	0.68	0.56	0.62	170
Accuracy	0.60			
Macro Avg	0.63	0.61	0.60	507
Weighted Avg	0.65	0.60	0.61	507

(d) Subtask D (Humor Detection)

Class (ID)	Precision	Recall	F1-score	Support
No Humor (0)	0.61	0.68	0.65	165
Humor (1)	0.84	0.79	0.82	342
Accuracy	0.76			
Macro Avg	0.73	0.74	0.73	507
Weighted Avg	0.77	0.76	0.76	507

in Table 3. As seen in the confusion matrix in Figure 5, the model often confuses Humor (1) with No Humor (0), misclassifying 71 humorous instances. This suggests the model is conservative in predicting humor, likely due to the subtle and context-dependent nature of humor in online content.

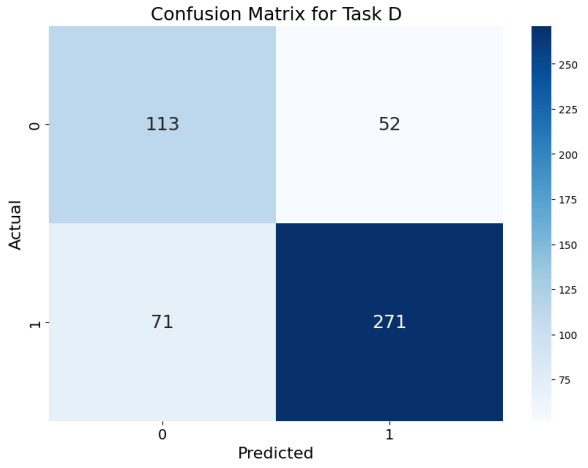

Figure 5: Confusion Matrix for SubTask-D.

4.3 Limitations and Future Enhancements

While the models demonstrate solid performance, several challenges remain. The confusion matrices highlight difficulty in distinguishing semantically similar or nuanced classes, reflecting the limits of current embedding and classification approaches in capturing subtle context, sarcasm, or implicit meanings. Although our approach used multimodal signals via CLIP, improvements could come from better integration techniques that more effectively fuse text and image information. Incorporating large language models (LLMs) for generating predictions or augmenting data could enhance understanding of complex language patterns and improve classification accuracy (Thapa et al., 2025b). Additionally, experimenting with larger pretrained models or ensembling strategies may boost robustness. Exploring advanced data augmentation or synthetic data generation to address class imbalance and rare cases could also enhance performance. Finally, incorporating domain-specific knowledge or interpretability techniques would help understand and mitigate systematic biases and errors.

5 Conclusion

In this work, we presented a unified multimodal framework based on the CLIP model to address multiple subtasks related to hate speech, target classification, stance detection, and humor detection. Our approach demonstrates strong performance across these classification challenges, effectively leveraging both textual and visual information. While results indicate potential, especially for hate speech and humor detection, challenges remain in handling subtle distinctions and class imbalances. Future improvements may involve deeper integration of multimodal cues and the use of large language models to better capture context and nuance. Overall, this study contributes to advancing robust, multimodal methods for understanding complex social content in online platforms.

References

Pamela Aronson and Islam Jaffal. 2021. Zoom memes for self-quaranteens: Generational humor, identity, and conflict during the pandemic. *Emerging Adulthood*.

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

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.

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)*.

Charunayan Kamath and Sivakumar Alur. 2024. Research trends in memes: Insights from bibliometric analysis. *Information Discovery and Delivery*.

Anil Singh Parihar, Surendrabikram Thapa, and Sushruti Mishra. 2021. Hate speech detection using natural language processing: Applications and challenges. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1302–1308. IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

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.