Multimodal Hate, Humor, and Stance Event Detection in Marginalized Sociopolitical Movements

Surendrabikram Thapa[1], **Siddhant Bikram Shah**[2], **Kritesh Rauniyar**[3, 4], **Shuvam Shiwakoti**[1], **Surabhi Adhikari**[5], **Hariram Veeramani**[6], **Kristina T. Johnson**[2], **Ali Hürriyetoğlu**[7], **Hristo Tanev**[8], **Usman Naseem**[9]

[1]Virginia Tech, USA, [2]Northeastern University, USA,
[3]Delhi Technological University, India, [4]IIMS College, Nepal, [5]Columbia University, USA,
[6]UCLA, USA, [7]Wageningen Food Safety Research, Netherlands,
[8]European Commission, Joint Research Centre, Italy, [9]Macquarie University, Australia
[1]{surendrabikram, shuvam}@vt.edu, [3]rauniyark11@gmail.com,
[7]ali.hurriyetoglu@wur.nl, [8]hristo.tanev@ec.europa.eu

Abstract

This paper presents the Shared Task on Multimodal Detection of Hate Speech, Humor, and Stance in Marginalized Socio-Political Movement Discourse, hosted at CASE 2025. The task is built on the PrideMM dataset, a curated collection of 5,063 text-embedded images related to the LGBTQ+ pride movement, annotated for four interrelated subtasks: (A) Hate Speech Detection, (B) Hate Target Classification, (C) Topical Stance Classification, and (D) Intended Humor Detection. Eighty-nine teams registered, with competitive submissions across all subtasks. The results show that multimodal approaches consistently outperform unimodal baselines, particularly for hate speech detection, while fine-grained tasks such as target identification and stance classification remain challenging due to label imbalance, multimodal ambiguity, and implicit or culturally specific content. CLIP-based models and parameter-efficient fusion architectures achieved strong performance, showing promising directions for low-resource and efficient multimodal systems.

1 Introduction

In the ever-evolving digital landscape, social media has become a pivotal arena for discourse, particularly for marginalized socio-political movements (Bhandari et al., 2023; Shiwakoti et al., 2024). Within these online spaces, text-embedded images, like memes, have emerged as a powerful and prevalent medium of communication. They serve as potent vehicles for expressing solidarity, fostering resistance, and shaping attitudes and perceptions both within and beyond these communities. The multimodal nature of memes, combining imagery and text, presents a formidable challenge for machine learning systems, which must move beyond simplistic analyses to grasp the multifaceted expressions conveyed (Pramanick et al., 2021b). As platforms increasingly grapple with content moderation challenges, the ambiguity between satire and offense in such imagery underscores a critical gap in computational analysis: multimodal understanding must disentangle layered communicative intents to mitigate harm while preserving cultural context (Scott, 2021).

The discourse surrounding marginalized communities is often complex and multifaceted, where the lines between humor, satire, and genuine harm are frequently blurred (Klassen and Fiesler, 2022). Memes, in this context, can simultaneously be instruments of empowerment and weapons of oppression, making the task of content moderation exceptionally difficult. A single label often fails to capture the layered meanings embedded within these images. Consequently, there is a pressing need for a more nuanced, multi-aspect understanding of such content to develop more effective AI systems (Pramanick et al., 2021a). The PrideMM dataset epitomizes this complexity, centering on discourse surrounding the LGBTQ+ movement where memes frequently blur the lines between humor and harm (Shah et al., 2024).

To address this critical research gap, building on our previous shared tasks at CASE 2024 (Thapa et al., 2024b; Hürriyetoğlu et al., 2024) and CASE 2023 (Thapa et al., 2023a; Hürriyetoğlu et al., 2023) we present the Shared Task on Multimodal Detection of Hate Speech, Humor, and Stance in Marginalized Socio-Political Movement Discourse at CASE 2025 (Hürriyetoğlu et al., 2025). This task utilizes the PrideMM dataset, a curated collection of memes related to the LGBTQ+ pride movement. The shared task is designed to spur the development of models that can analyze text-embedded images from four distinct yet interconnected perspectives: 1) Detection of Hate Speech, 2) Classifying the

Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025, pages 20–31, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-003

Targets of Hate Speech, 3) Classification of Topical Stance, and 4) Detection of Intended Humor. We frame these subtasks together to encourage holistic approaches that can capture the entangled social, cultural, and affective dimensions of online content.

This paper provides a comprehensive overview of the shared task, including a description of the dataset, the evaluation metrics for each subtask, a summary of the participating teams and their methodologies, and an analysis of the results. Through this shared task, we aim to foster innovation in multimodal analysis and contribute to the development of more sophisticated and context-aware models for understanding online discourse.

2 Dataset

For our shared task, we utilize the PrideMM dataset, as shown in Table 1, introduced by Shah et al. (2024), which comprises a total of 5,063 text-embedded images (memes, posters, and infographics) related to the LGBTQ+ Pride movement. This multimodal dataset addresses the need for more inclusive and nuanced resources in the meme analysis space by encompassing four distinct yet related tasks: (A) Hate Speech Detection, (B) Hate Target Classification, (C) Topical Stance Classification, and (D) Humor Detection. The dataset spans content from 2020 to 2024, collected from Twitter, Facebook, and Reddit using targeted queries and community-specific groups. To ensure high-quality and diverse samples, redundant images were filtered using deduplication tools, and OCR was applied to extract and clean embedded text. Each image in the dataset was independently annotated by five trained annotators through a rigorous three-phase annotation protocol to enhance label consistency. The annotators labeled every image for hate presence, stance, and humor, and for hateful images, also annotated the target (undirected, individual, community, or organization).

3 Shared Task Description

Social media platforms amplify controversial content and quickly disseminate conflict. Meanwhile, humorous content, especially memes, has become more popular as a tool for digital community involvement and influence (Pramanick et al., 2021a). As people become more outspoken about their opinions, stance identification is essential when analyzing public opinion (AlDayel and Magdy, 2021). This shared task focuses on investigating the potential of different multimodal models and the identification of four unique tasks related to socio-political discourse. Further details on subtasks can be found below:

Subtask	Classes	Train	Eval	Test	Total
Subtask A	Hate	1,985	248	249	2,482
	No-Hate	2,065	258	258	2,581
Subtask B	Undirected	617	77	77	771
	Individual	199	25	25	249
	Community	931	116	117	1,164
	Organization	238	30	30	298
Subtask C	Neutral	1,166	146	146	1,458
	Support	1,527	191	191	1,909
	Oppose	1,357	169	170	1,696
Subtask D	No Humor	1,313	164	165	1,642
	Humor	2,737	342	342	3,421

Table 1: Statistics of the dataset provided to the participants as part of the shared task.

3.1 Subtask A: Hate Speech Detection

The primary objective of this subtask is to identify the presence of hate speech in the text-embedded images. This is a binary classification task, where each sample is annotated with one of two possible labels: *Hate Speech* or *No Hate Speech*. The dataset used for this task primarily concentrates on visuals in which text is important to expressing meaning, facilitating a more nuanced analysis of harmful or offensive content. Using both textual and visual, the dataset offers clear separation between the two categories.

3.2 Subtask B: Targets of Hate Speech Detection

For content identified as hateful, this subtask requires a more granular analysis to determine the target of the hate speech in the images. The image contains hateful text, with the determined target categories as *Individual*, *Community*, *Organization*, or *Undirected*. This subtask focuses on four specific categories within the text-embedded images, which help to identify and understand the type of hateful content.

3.3 Subtask C: Topical Stance Detection

This subtask aims to classify the stance of the meme towards the marginalized movement itself, with possible labels of *Support*, *Oppose*, or *Neutral*. This subtask involves classifying and understanding the stance of the meme images, with a focus on understanding the type of stance that facilitates grasping the categories of stance.

3.4 Subtask D: Intended Humor Detection

Recognizing the prevalence of satire and humor in this domain, this subtask challenges participants to identify whether a meme is intended to be humorous, using *Humor* or *No Humor* labels.

4 Participants' Methods

4.1 Overview

Of the 89 registered participants, 21 submitted results for Subtask A, 14 for Subtask B, 13 for Subtask C, and 16 for Subtask D. The leaderboards for these subtasks are presented in Table 2, table 3, table 4 and table 5.

4.2 Methods

The following section presents brief overviews of the participating teams' approaches, based on the methodologies outlined in their system description papers.

4.2.1 Subtask A: Hate Speech Detection

TSR (Ray et al., 2025) presented FIMIF (Feature Interaction for Multi-Modal Integration and Fusion), a lightweight multimodal framework that relies on frozen CLIP ViT-L/14 encoders for extracting text and image embeddings. These embeddings were compressed into low-dimensional spaces using residual projection layers before being passed to a multiplicative feature interaction module designed to capture higher-order cross-modal relationships. Their approach emphasizes efficiency, with only 25k–51k trainable parameters, yet it achieved an F1-score of 81.85% and accuracy of 81.85% in hate speech detection. This demonstrates that dimensionality compression coupled with multiplicative fusion can yield competitive results on multimodal hate classification.

PhantomTroupe (Amin et al., 2025) experimented with multiple approach, including unimodal and multimodal, where the fine-tuned Qwen2.5-VL-7B-Instruct- bnb-4bit using the unsloth framework outperformed acheiving the F1-score of 80.86%. Both approaches followed a transformer-based model, placing the team in 5th position.

MemeMasters (Shakya and Gurung, 2025) utilized a fine-tuned CLIP model as their primary architecture. Their approach involved concatenating the visual and textual embeddings from the CLIP model and feeding them into a lightweight classification head. Using their standard configuration without task-specific modifications, the system achieved a macro F1-score of 80%, showing consistent and balanced performance across both the "Hate" and "Non-Hate" classes.

Multimodal Kathmandu (Maharjan et al., 2025) employed a Co-Attention Ensemble architecture built upon frozen CLIP-ViT features. Text and image embeddings were concatenated and passed through multi-layer Transformer encoders, with predictions averaged across five ensemble members to reduce variance. This approach achieved an F1-score of 79.29% and an accuracy of 79.29%, highlighting the robustness of variance-reduction strategies for multimodal hate speech detection.

MLInitiative (Acharya et al., 2025) investigated two multimodal architectures, a ResNet-18 with BERT model and a SigLIP2 model, for hate speech detection. Their fine-tuned SigLIP2 model outperformed the ResNet-18+BERT baseline, achieving an F1-score of 79.27%. This performance placed their system 9th on the final leaderboard for the subtask.

ID4Fusion (Rashfi et al., 2025) utilized transformer-based models, RoBERTa and HateBERT were fine-tuned for text analysis, while EfficientNet-B7 and Vision Transformer (ViT) were utilized for images. The predictions from these models were integrated using a late-fusion ensemble approach, providing more weight to textual features compared to visual features. In the leaderboard, they secured 10th position.

Silver (Mainali et al., 2025) evaluated a range of unimodal and multimodal models, including transformer-based text models like BERT and ROBERTa, CNN-based vision models like DenseNet and EfficientNet, and fusion methods like CLIP. Their results demonstrated that multimodal systems performed better than unimodal baselines, with a CLIP-based model achieving the top macro F1-score of 78.28%. The authors noted that models often misclassified sarcastic or ironic content where hate was conveyed visually rather than through explicit text.

Rank	Team Name	Codalab Username	F1-score (%)	Accuracy (%)	Precision (%)	Recall (%)
1	TJU-MI	wangxiuxian	84.22	84.22	84.22	84.22
2	-	Ryuan	82.84	82.84	82.91	82.88
3	IMU-L	jiaranDiana	82.05	82.05	82.17	82.11
4	TSR (Ray et al., 2025)	ray-sushant	81.85	81.85	81.91	81.89
5	Phantom Troupe (Amin et al., 2025)	Neuron-Force	80.86	80.87	80.86	80.86
6	MemeMasters (Shakya and Gurung, 2025)	shrutigurung	80.05	80.08	80.12	80.04
7	Multimodal Kathmandu (Maharjan et al., 2025)	Sujal_Maharjan	79.29	79.29	79.33	79.32
8	-	NextTry	79.28	79.29	79.28	79.28
9	MLInitiative (Acharya et al., 2025)	ankitbk07	79.27	79.29	79.30	79.27
10	ID4Fusion (Rashfi et al., 2025)	Rashfi	78.68	78.70	78.70	78.68
11	Silver (Mainali et al., 2025)	rohanmainali	78.28	78.30	78.33	78.27
12	MMFusion (Rane, 2025)	prerana3	77.89	77.91	78.14	77.99
13	CUET NOOB (Joy et al., 2025)	TomalJoy	74.16	74.16	74.16	74.17
14	-	Tanvir_77	74.02	74.16	74.47	74.05
15	Overfitters (Bhattarai et al., 2025)	bidhancb	73.77	73.77	73.82	73.80
16	-	AkshYat	73.37	73.37	73.47	73.42
17	Luminaries (Esackimuthu, 2025)	akshayy22	72.17	72.19	72.34	72.25
18	YS	ysb	69.23	69.23	69.27	69.26
19	MLP (Verma and Kumar, 2025)	Durgeshverma24iitram	66.02	66.27	66.54	66.14
20	wangkongqiang (Kongqiang and Peng, 2025)	wangkongqiang	62.09	63.31	65.91	63.65
21	Musafir	MDSagorChowdhury	58.28	62.33	68.62	61.79

Table 2: Sub-task A (Hate Speech Classification) Leaderboard, Ranked by Macro F1-score. All scores are presented as percentages (%). It is to be noted that this leaderboard contains the score till the test deadline and does not consider further runs done by participants as a part of the system description paper.

MMFusion (Rane, 2025) implemented a multimodal architecture using RoBERTa-base for textual analysis and a ResNet50 model for visual feature extraction. These features were projected into a shared dimensional space and combined using an 8-head multi-head attention mechanism to capture cross-modal interactions. The team also employed focal loss to concentrate on difficult samples and used a test-time augmentation (TTA) strategy to improve robustness, ultimately achieving an F1-score of 77.8%.

CUET NOOB (Joy et al., 2025) used the multimodal attention-based late fusion approach to capture cross-modal interactions. The model achieved an F1-score of 74.16%, ranking 13th overall. The authors also experimented with unimodal models like DistilBERT for text and ViT for images.

Overfitters (Bhattarai et al., 2025) utilized a multimodal fusion model named BERTRES for hate speech detection, combining textual features from a BERT-base model with visual embeddings from a ResNet-50 model. The concatenated feature vector was processed through a classifier with separate heads for each task. This approach achieved an F1-score of 73.77%, placing them 15th in the task. The paper suggests the model's performance was limited by the difficulty of capturing subtle, implied hate and sarcasm in memes.

Luminaries (Esackimuthu, 2025) explored a hybrid modeling approach by combining the ALBERT-base v2 transformer with classical machine learning models such as XGBoost, LightGBM, Gradient Boosting, and MLP classifiers. Predictions from these systems, trained on TF-IDF and syntactic features alongside contextual embeddings, were integrated through a weighted ensembling strategy. This ensemble achieved an F1-score of 72.17%, ranking 17th overall. The authors note that ensembling effectively leveraged complementary strengths across models, though additional linguistic features and further tuning of ensemble weights could yield improvements.

MLP (Verma and Kumar, 2025) developed multimodal frameworks that fused XLM-RoBERTa and ViT embeddings with attention-based fusion, as well as alternative combinations with CLIP and BERT encoders. Their best-performing configuration achieved an F1-score of 66.02% and an accuracy of 66.27%. The system demonstrated the effectiveness of early fusion and cross-modal attention in detecting hate content from memes.

wangkongqiang (Kongqiang and Peng, 2025) employed different approaches, including an ensemble model integrating text and image features (utilizing BERT, XLNet, and InceptionNet), a K-max pooling neural network utilizing pre-trained GloVe embeddings and cyclic learning rate scheduling, and a multinomial naive Bayes (MNB) model. The MNB achieved an F1-score of 62.09%, which placed them in 20th position.

Rank	Team Name	Codalab Username	F1-score (%)	Accuracy (%)	Precision (%)	Recall (%)
1	TJU-MI	wangxiuxian	65.30	64.26	67.59	63.83
2	-	Ryuan	63.35	64.26	65.56	62.04
3	TSR (Ray et al., 2025)	ray-sushant	60.57	63.05	61.69	60.21
4	IMU-L	jiaranDiana	60.15	63.05	62.30	60.38
5	Multimodal Kathmandu (Maharjan et al., 2025)	Sujal_Maharjan	57.77	58.23	56.66	59.22
6	Overfitters (Bhattarai et al., 2025)	bidhancb	56.28	57.03	54.07	60.32
7	MMFusion (Rane, 2025)	prerana3	55.39	59.04	56.53	55.04
8	MLInitiative (Acharya et al., 2025)	ankitbk07	54.86	58.23	60.44	52.49
9	MemeMasters (Shakya and Gurung, 2025)	shrutigurung	51.50	53.82	54.27	50.59
10	Silver (Mainali et al., 2025)	rohanmainali	50.18	51.81	50.92	54.22
11	Luminaries (Esackimuthu, 2025)	akshayy22	49.84	55.42	52.89	48.69
12	Musafir	MDSagorChowdhury	37.93	44.18	40.08	41.43
13	wangkongqiang (Kongqiang and Peng, 2025)	wangkongqiang	34.53	47.79	55.52	33.22
14	MLP (Verma and Kumar, 2025)	Durgeshverma24iitram	27.39	40.96	31.58	27.57

Table 3: Sub-task B (Target Identification for Hate Speech) Leaderboard, Ranked by Macro F1-score. All scores are presented as percentages (%). It is to be noted that this leaderboard contains the score till the test deadline and does not consider further runs done by participants as a part of the system description paper.

4.2.2 Subtask B: Targets of Hate Speech Detection

TSR (Ray et al., 2025) adapted their FIMIF pipeline by addressing the severe class imbalance in the dataset. They incorporated weighted cross-entropy loss and deterministic oversampling of minority classes to stabilize learning across the four target categories (Undirected, Individual, Community, Organization). The system combined CLIP-based representations through residual and multiplicative modules, reaching an F1-score of 60.57% and accuracy of 63.05%. The results indicate that their compact architecture could model nuanced target categories effectively despite relying on a parameter-efficient design.

Multimodal Kathmandu (Maharjan et al., 2025) designed a Hierarchical Cross-Attention Transformer that allowed textual tokens to query visual regions directly. This task-specific architecture reached an F1-score of 57.77% with an accuracy of 58.23%, ranking 5th on the leaderboard. The results underscore the value of explicit cross-modal grounding for distinguishing between Community, Individual, Organization, and Undirected targets.

Overfitters (Bhattarai et al., 2025) applied their BERTRES architecture, which fuses features from BERT and ResNet-50. A key part of their methodology was the use of a class-weighted cross-entropy loss to mitigate the skewed label distribution present in the dataset for this subtask. This strategy contributed to their strongest performance, securing the 6th position with an F1-score of 56.28%.

MMFusion (Rane, 2025) utilized their RoBERTa-ResNet50 architecture with cross-modal attention for the target identification task. To address the significant class imbalance in this subtask, they used a focal loss function with class-specific weighting in addition to a test-time augmentation strategy. The model struggled to differentiate between certain categories, particularly confusing ”Individual” targets with ”Community” targets, and achieved a final F1-score of 55.3%.

MLInitiative (Acharya et al., 2025) applied two multimodal systems: a combined ResNet-18 and BERT architecture and a SigLIP2 model. The SigLIP2 model proved to be superior, securing an F1-score of 54.86% and ranking 8th on the leaderboard. The authors note that both models performed relatively poorly on this task, attributing the difficulty to the imbalanced nature of the associated dataset.

MemeMasters (Shakya and Gurung, 2025) adapted their fine-tuned CLIP model for the target classification task by applying over-sampling to mitigate the severe class imbalance present in the dataset. The model struggled with the fine-grained nature of this task, showing the lowest recall for the ”Individual” class and frequent confusion between the ”Undirected” and ”Community” targets. This resulted in a macro F1-score of 52%.

Silver (Mainali et al., 2025) addressed the target classification task, noting it was particularly challenging due to the highly uneven distribution of classes and the subjective nature of defining a target. Comparing various unimodal and multimodal systems, their CLIP-based model again achieved the best performance with a macro F1-score of 56.30%. This score represented a considerable performance decline compared to other subtasks, with the paper highlighting

Rank	Team Name	Codalab Username	F1-score (%)	Accuracy (%)	Precision (%)	Recall (%)
1	TJU-MI	wangxiuxian	63.70	64.89	64.59	63.68
2	TSR (Ray et al., 2025)	ray-sushant	62.91	62.92	64.22	63.07
3		Ryuan	62.80	62.72	64.50	63.10
4	IMU-L	jiaranDiana	61.76	62.13	62.54	61.58
5	MMFusion (Rane, 2025)	prerana3	60.86	61.14	61.23	61.14
6	Multimodal Kathmandu (Maharjan et al., 2025)	Sujal_Maharjan	60.70	61.14	61.18	61.25
7	MLInitiative (Acharya et al., 2025)	ankitbk07	60.59	61.14	60.64	60.59
8	MemeMasters (Shakya and Gurung, 2025)	shrutigurung	60.23	59.96	63.09	60.56
9	Overfitters (Bhattarai et al., 2025)	bidhancb	60.15	60.55	60.27	60.17
10	Silver (Mainali et al., 2025)	rohanmainali	59.30	59.57	59.53	59.47
11	Musafir	MDSagorChowdhury	54.29	54.24	55.55	54.83
12	Luminaries (Esackimuthu, 2025)	akshayyy22	53.05	55.23	54.34	53.55
13	MLP (Verma and Kumar, 2025)	Durgeshverma24iitram	46.74	46.94	49.05	47.23

Table 4: Sub-task C (Classification of Topical Stance) Leaderboard, Ranked by Macro F1-score. All scores are presented as percentages (%). It is to be noted that this leaderboard contains the score till the test deadline and does not consider further runs done by participants as a part of the system description paper.

major confusion between the ’Community’ and ’Undirected’ classes as a key issue.

Luminaries (Esackimuthu, 2025)fine-tuned ALBERT for multiclass classification and also trained a feedforward ANN. Their ALBERT model achieved an F1-score of 0.4984 and an accuracy of 55.42%, placing 11th on the leaderboard. While the model effectively captured contextual dependencies, the system struggled to distinguish between fine-grained target categories.

wangkongqiang (Kongqiang and Peng, 2025) performed four different benchmarks in different models, where the multinomial naive bayes classification model showed the best, achieving an F1-score of 34.53% and holding 13th position in the leaderboard.

MLP (Verma and Kumar, 2025) employed XLM-RoBERTa + ViT with attention-based fusion. The system achieved an F1-score of 40.96% and an accuracy of 42.17%. Despite leveraging bidirectional cross-modal attention and multiple fusion strategies, the model struggled with fine-grained classification of targets within memes.

4.2.3 Subtask C: Topical Stance Detection

TSR (Ray et al., 2025) applied the FIMIF architecture with low-rank multimodal fusion of compressed text and image embeddings. The model obtained an F1-score of 62.91% and accuracy of 62.92%. The architecture leveraged CLIP embeddings alongside residual layers that emphasized linear relationships, with the multiplicative module capturing more complex feature interactions when beneficial. These results highlight the model’s ability to distinguish Support, Oppose, and Neutral stances in multimodal memes with relatively few parameters.

MMFusion (Rane, 2025) adopted an ensemble approach, combining the outputs of three separate multimodal models via probability averaging. The ensemble consisted of a RoBERTa-base with ResNet18, a RoBERTa-base with ResNet34, and a BERT-base with ResNet18, each trained with different random seeds to ensure diversity. This method, which performed better than their initial single-model attempts, used a simple attention mechanism for feature fusion and yielded an F1-score of 60.8%.

Multimodal Kathmandu(Maharjan et al., 2025) introduced a Two-Stage Multiplicative Fusion framework, where CLIP features were projected into higher dimensions, refined through lightweight adapters, and combined via element-wise multiplication. A two-stage fine-tuning procedure stabilized training and improved convergence. Their model achieved an F1-score of 60.70% and accuracy of 61.14%, placing 6th overall.

MLInitiative (Acharya et al., 2025) compared their ResNet-18+BERT and SigLIP2 multimodal models. The SigLIP2 model, which processes image-text pairs in a joint embedding space using a sigmoid-based contrastive loss, obtained better performance. It achieved an F1-score of 60.59%, which resulted in a 7th place ranking on the task leaderboard.

MemeMasters (Shakya and Gurung, 2025) modified their CLIP-based framework by employing a deeper 3-layer classifier head and using a cosine learning rate scheduler. The model found it difficult to distinguish between subtle stance dif-

Rank	Team Name	Codalab Username	F1-score (%)	Accuracy (%)	Precision (%)	Recall (%)
1	TJU-MI	wangxiuxian	78.01	81.07	77.50	78.65
2	TSR (Ray et al., 2025)	ray-sushant	76.83	79.68	76.79	76.87
3	Multimodal Kathmandu (Maharjan et al., 2025)	Sujal_Maharjan	75.29	77.91	75.78	74.91
4	-	paiisfunny	75.16	78.50	74.81	75.57
5	-	Ryuan	74.80	78.30	74.35	75.37
6	MemeMasters (Shakya and Gurung, 2025)	shrutigurung	73.13	75.74	73.86	72.66
7	MLInitiative (Acharya et al., 2025)	ankitbk07	72.88	77.71	71.72	75.26
8	Silver (Mainali et al., 2025)	rohanmainali	72.68	75.94	72.75	72.61
9	-	olivialiudama	71.41	75.35	71.06	71.86
10	-	AkshYat	71.13	73.18	72.75	70.67
11	IMU-L	jiaranDiana	70.31	72.98	71.19	69.85
12	MMFusion (Rane, 2025)	prerana3	65.85	73.37	64.89	70.05
13	MLP (Verma and Kumar, 2025)	Durgeshverma24iitram	65.64	70.02	65.54	65.75
14	Overfitters (Bhattarai et al., 2025)	bidhancb	65.33	72.78	64.46	69.06
15	Musafir	MDSagorChowdhury	62.68	66.07	63.25	62.44
16	Luminaries (Esackimuthu, 2025)	akshayyy22	60.70	68.44	60.30	62.74

Table 5: Sub-task D (Classification of Intended Humor) Leaderboard, Ranked by Macro F1-score. All scores are presented as percentages (%). It is to be noted that this leaderboard contains the score till the test deadline and does not consider further runs done by participants as a part of the system description paper.

ferences, leading to significant misclassifications where both 'Support' and 'Oppose' instances were incorrectly labeled as 'Neutral'. This approach achieved a macro F1-score of 60%.

Overfitters (Bhattarai et al., 2025) implemented their BERTRES model, which leverages a combination of BERT and ResNet-50 embeddings. The system struggled with the complexities of this task, particularly in cases that involved satire or ambiguous sentiment. The model also demonstrated overfitting, which the authors attribute to the imbalanced and sparse label distribution for this specific subtask. Ultimately, the system ranked 9th with an F1-score of 60.15%.

Silver (Mainali et al., 2025) employed their comparative framework of unimodal and multimodal models. Their best system was a CLIP-based model, which achieved a macro F1-score of 59.30%. Even with this top-performing model, the authors reported that it struggled to correctly classify memes containing irony or ambiguous tone. The 'Neutral' class was the most likely to be miscategorized as either supportive or opposing.

Luminaries (Esackimuthu, 2025) trained ALBERT and a feedforward ANN independently for stance classification without ensembling. The ALBERT model achieved an F1-score of 53.05%, ranking 12th overall. The system faced difficulty in interpreting ambiguous or ironic stances, which often rely on subtle linguistic cues.

MLP (Verma and Kumar, 2025) applied their multimodal attention-based fusion models. The best-performing system achieved an F1-score of 46.74% with an accuracy of 46.94%. While the approach captured some multimodal interactions, performance remained limited on subtle stance distinctions.

4.2.4 Subtask D: Intended Humor Detection

TSR (Ray et al., 2025) achieved their strongest results, reporting an F1-score of 76.83% and accuracy of 79.68%. The FIMIF model effectively integrated textual sarcasm with visual cues through its multiplicative fusion layer, enabling the detection of humor and satire in memes. Despite its small size compared to transformer-based multimodal systems, the architecture maintained competitive performance, underscoring the value of low-dimensional fusion for subjective classification tasks.

Multimodal Kathmandu(Maharjan et al., 2025) applied the same Two-Stage Multiplicative Fusion framework, augmented with semantic-aware initialization that seeded classifier weights using CLIP embeddings of descriptive prompts. This system delivered one of the strongest results in the task, achieving an F1-score of 75.29% and accuracy of 77.91%, securing 3rd place overall.

MemeMasters (Shakya and Gurung, 2025) adjusted their CLIP model by incorporating a higher dropout rate and using a class-weighted loss to handle the imbalanced data, which was skewed towards humorous content. The model was conservative in its predictions, often misclassifying humorous content as non-humorous due to the context-dependent nature of online humor. This system yielded a macro F1-score of 73%.

MLInitiative (Acharya et al., 2025) addressed humor detection using a ResNet-18+BERT fusion model and a more advanced SigLIP2 model. Their results showed that the SigLIP2 architecture was more effective for the task, achieving an F1-score of 72.88%. This performance earned their system the 7th position on the subtask's final leaderboard.

Silver (Mainali et al., 2025) found that multimodal models outperformed unimodal approaches, as visual cues were often critical for contextualizing humor in memes. A CLIP-based model proved to be the most effective, delivering a macro F1-score of 72.68%. Despite this success, the system was prone to making false predictions on content that involved sarcasm or culturally specific jokes that were not conveyed through text.

MMFusion (Rane, 2025) developed a distinct multimodal architecture using DialoGPT-medium for text and ResNet50 for images, choosing DialoGPT for its proficiency with informal, conversational language. Their system applied self-attention to each modality independently before using cross-modal attention and a final gating mechanism to adaptively weight and combine the features. This approach resulted in an F1-score of 65.8%.

MLP (Verma and Kumar, 2025) reported stronger results relative to stance and target identification. Their multimodal architecture reached an F1-score of 65.64% and an accuracy of 70.02%. This indicates that their system was able to capture explicit humorous cues in memes using cross-modal fusion of textual and visual features.

Overfitters (Bhattarai et al., 2025) addressed the humor detection challenge with their BERTRES multimodal fusion model. The model, which integrates text embeddings from BERT and visual features from ResNet-50, found this task to be particularly difficult due to the subjective and culturally specific nature of humor in memes, which made it hard for the model to generalize. Their system achieved an F1-score of 65.33%, resulting in a 14th-place ranking.

Luminaries (Esackimuthu, 2025) utilized ALBERT and an ANN, treating this as a binary classification task. Their fine-tuned ALBERT model achieved an F1-score of 60.70%, ranking 16th overall. Performance was constrained by the subjective and culturally dependent nature of humor, with frequent misclassification of sarcastic or context-heavy instances.

5 Discussion

The results across the four subtasks show the complexities of multimodal analysis of socio-political memes, where humor, satire, and harmful speech often intersect. While multimodal models generally outperformed unimodal baselines, the extent of improvement varied by task, reflecting differences in difficulty, class imbalance, and the interplay of textual and visual cues. Hate Speech Detection (Subtask A) achieved the highest scores, with several teams surpassing an F1-score of 80%, indicating that binary classification of explicit or strongly implied hate is relatively well-handled by current models. In contrast, Target Identification (Subtask B) proved most challenging, with substantial performance drops due to fine-grained labels, skewed class distributions, and frequent overlap between categories such as Community and Undirected.

Stance Detection (Subtask C) showed moderate performance, with top scores in the low 60s, hindered by the difficulty of interpreting sarcasm, irony, and ambiguous sentiment. Humor Detection (Subtask D) fared slightly better, with top teams exceeding 76% F1, suggesting that visual tropes and textual patterns characteristic of humor are more consistently captured by multimodal fusion methods. CLIP-based approaches dominated many leaderboards, while compact, parameter-efficient architectures like TSR's FIMIF (Ray et al., 2025) demonstrated that strong results are achievable with minimal trainable parameters. Attention-based and gating fusion mechanisms yielded mixed benefits, with improvements often dependent on task-specific dynamics.

Persistent challenges include handling subtle, culturally dependent cues, mitigating severe class imbalance, particularly in Subtask B, and resolving multimodal ambiguity when text and image signals conflict or provide weak cues. Future progress will require better handling of fine-grained categories, integration of external knowledge to interpret implicit references, improved balancing strategies, and hybrid architectures that combine precise language understanding with strong cross-modal alignment. Overall, while current models show promise in detecting overt hate and humor, capturing nuanced communicative intent in marginalized community discourse remains an open challenge.

6 Conclusion

In this paper, we presented the Shared Task on Multimodal Detection of Hate Speech, Humor, and Stance in Marginalized Socio-Political Movement Discourse at CASE 2025 (co-located with RANLP 2025), leveraging the PrideMM dataset, a curated collection of memes related to the LGBTQ+ pride movement. The task was designed to encourage the development of models capable of jointly addressing four interconnected challenges: (i) detecting hate speech, (ii) identifying its targets, (iii) classifying topical stance, and (iv) recognizing intended humor. With participation from 89 registered teams and competitive submissions across all subtasks, the results demonstrated the clear advantage of multimodal approaches over unimodal baselines, while also revealing substantial variation in task difficulty and persistent challenges in handling subtle, implicit, or culturally dependent content. The insights from this shared task provide a valuable benchmark for future research, suggesting the need for methods that combine robust cross-modal integration with cultural and contextual awareness. We hope this work will stimulate continued innovation in multimodal content moderation, contributing to safer and more inclusive online spaces.

Acknowledgments

Funding for part of this research has been provided by the European Union's Horizon Europe research and innovation programme EFRA [grant number 101093026] and ECO-Ready [grant number 101084201].

Broader Impact

This shared task aims to advance multimodal content moderation in contexts involving marginalized socio-political movements, focusing on the LGBTQ+ pride movement. By targeting nuanced, culturally embedded, and often ambiguous content, it encourages the development of fairer, more context-aware AI systems that can mitigate harm while preserving legitimate expression. However, automated moderation must be applied cautiously, as misclassification can silence marginalized voices or misinterpret culturally specific discourse. The PrideMM dataset was curated with rigorous annotation to promote inclusivity and reduce bias, but real-world use should involve human oversight and community input. Beyond moderation, the work offers value for social science, media studies, and policy, supporting safer and more inclusive online spaces while respecting expressive diversity.

References

Ashish Acharya, Ankit BK, Bikram K.C., Sandesh Shrestha, Tina Lama, Surabhi Adhikari, and Rabin Thapa. 2025. MLInitiative at CASE 2025: Multimodal Detection of Hate Speech, Humor,and Stance using Transformers. In *Proceedings of The Eighth Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Abeer AlDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Farhan Amin, Muhammad Abu Horaira, Md. Tanvir Ahammed Shawon, Md. Ayon Mia, and Muhammad Ibrahim Khan. 2025. PhantomTroupe@CASE 2025: Multimodal Hate Speech Detection in Text-Embedded Memes using Instruction-Tuned LLMs. In *Proceedings of The Eighth Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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.

Bidhan Chandra Bhattarai, Dipshan Pokhrel, Ishan Maharjan, and Rabin Thapa. 2025. Overfitters@CASE2025: Multimodal Hate Speech Analysis Using BERT and RESNET. In *Proceedings of The Eighth Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Jamell Dacon, Harry Shomer, Shaylynn Crum-Dacon, and Jiliang Tang. 2022. Detecting harmful online conversational content towards lgbtqia+ individuals. *arXiv preprint arXiv:2207.10032*.

Akshay Esackimuthu. 2025. Luminaries@CASE 2025: Multimodal Hate Speech, Target, Stance and Humor Detection using ALBERT and Classical Models. In *Proceedings of The Eighth Workshop on Challenges and Applications of Automated Extraction of*

Socio-political Events from Text (CASE), Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Akash Gautam, Puneet Mathur, Rakesh Gosangi, Debanjan Mahata, Ramit Sawhney, and Rajiv Ratn Shah. 2020. # metooma: Multi-aspect annotations of tweets related to the metoo movement. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 209–216.

Ali Hürriyetoğlu, Hristo Tanev, Osman Mutlu, Surendrabikram Thapa, Fiona Anting Tan, and Erdem Yörük. 2023. Challenges and applications of automated extraction of socio-political events from text (CASE 2023): Workshop and shared task report. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text*, pages 167–175, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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)*.

Ali Hürriyetoğlu, Surendrabikram Thapa, Gökçe Uludoğan, Somaiyeh Dehghan, and Hristo Tanev. 2024. A concise report of the 7th workshop on challenges and applications of automated extraction of socio-political events from text. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 248–255.

Tomal Paul Joy, Aminul Islam, Saimum Islam, Md. Tanvir Ahammed Shawon, Md. Ayon Mia, and Mohammad Ibrahim Khan. 2025. CUET NOOB@CASE2025: MultimodalHate Speech Detection in Text-Embedded Memes using Late Fusion with Attention Mechanism. In *Proceedings of The Eighth Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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.

Shamika Klassen and Casey Fiesler. 2022. "this isn't your data, friend": Black twitter as a case study on research ethics for public data. *Social Media+ Society*, 8(4):20563051221144317.

Wang Kongqiang and Zhang Peng. 2025. wangkongqiang@CASE 2025: Detection and Classifying Language and Targets of Hate Speech using Auxiliary Text Supervised Learning. In *Proceedings of The Eighth Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Sujal Maharjan, Astha Shrestha, Shuvam Thakur, and Rabin Thapa. 2025. Multimodal Kathmandu@CASE 2025: Task-Specific Adaptation of Multimodal Transformers for Hate, Stance, and Humor Detection. In *Proceedings of The Eighth Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Rohan Mainali, Neha Aryal, Sweta Poudel, Anupraj Acharya, and Rabin Thapa. 2025. Silver@CASE2025: Detection of Hate Speech, Targets, Humor, and Stance in Marginalized Movement. In *Proceedings of The Eighth Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Usman Naseem, Shuvam Shiwakoti, Siddhant Bikram Shah, Surendrabikram Thapa, and Qi Zhang. 2025. Gametox: A comprehensive dataset and analysis for enhanced toxicity detection in online gaming communities. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 440–447.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. *arXiv preprint arXiv:1908.11049*.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021a. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796, Online. Association for Computational Linguistics.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021b. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jingnong Qu, Liunian Harold Li, Jieyu Zhao, Sunipa Dev, and Kai-Wei Chang. 2022. Disinfomeme: A multimodal dataset for detecting meme intentionally spreading out disinformation. *arXiv preprint arXiv:2205.12617*.

Prerana Rane. 2025. MMFusion@CASE 2025: Multimodal Learning for Hate Speech, Target, Stance, and Humor Classification in Marginalized Movement Discourse. In *Proceedings of The Eighth Workshop* *on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Tabassum Basher Rashfi, Md. Tanvir Ahammed Shawon, Md. Ayon Mia, and Muhammad Ibrahim Khan. 2025. ID4Fusion@CASE 2025: A Multimodal Approach to Hate Speech Detection in Text-Embedded Memes Using ensemble Transformer based approach. In *Proceedings of The Eighth Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Kritesh Rauniyar, Sweta Poudel, Shuvam Shiwakoti, Surendrabikram Thapa, Junaid Rashid, Jungeun Kim, Muhammad Imran, and Usman Naseem. 2023. Multi-aspect annotation and analysis of nepali tweets on anti-establishment election discourse. *IEEE Access*, 11:143092–143115.

Sushant Kr. Ray, Rafiq Ali, and Abdullah Mohammad. 2025. TSR@CASE 2025: Low Dimensional Multimodal Fusion Using Multiplicative Fine Tuning Modules. In *Proceedings of The Eighth Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Kate Scott. 2021. Memes as multimodal metaphors: A relevance theory analysis. *Pragmatics & Cognition*, 28(2):277–298.

Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. 2024. Memeclip: Leveraging clip representations for multimodal meme classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17320–17332, Miami, Florida, USA. Association for Computational Linguistics.

Shubham Shakya and Shruti Gurung. 2025. Team MemeMasters@CASE 2025: Adapting Vision-Language Models for Understanding Hate Speech in Multimodal Content. In *Proceedings of The Eighth Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Shuvam Shiwakoti, Surendrabikram Thapa, Kritesh Rauniyar, Akshyat Shah, Aashish Bhandari, and Usman Naseem. 2024. Analyzing the dynamics of climate change discourse on twitter: A new annotated corpus and multi-aspect classification. In *Joint 30th International Conference on Computational Linguistics and 14th International Conference on Language Resources and Evaluation, LREC-COLING 2024*, pages 984–994. European Language Resources Association (ELRA).

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 32–41.

Kohtaro Tanaka, Hiroaki Yamane, Yusuke Mori, Yusuke Mukuta, and Tatsuya Harada. 2022. Learning to evaluate humor in memes based on the incongruity theory. In *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, pages 81–93.

Surendrabikram Thapa, Farhan Ahmad Jafri, Ali Hürriyetoğlu, Francielle Vargas, Roy Ka Wei Lee, and Usman Naseem. 2023a. 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, Shuvam Shiwakoti, Hariram Veeramani, Raghav Jain, Guneet Singh Kohli, Ali Hürriyetoğlu, and Usman Naseem. 2024a. Stance and hate event detection in tweets related to climate activism-shared task at case 2024. In *7th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text, CASE 2024*, pages 234–247. Association for Computational Linguistics.

Surendrabikram Thapa, Kritesh Rauniyar, Farhan Ahmad Jafri, Hariram Veeramani, Raghav Jain, Sandesh Jain, Francielle Vargas, Ali Hürriyetoğlu, and Usman Naseem. 2024b. 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, Kritesh Rauniyar, Shuvam Shiwakoti, Sweta Poudel, Usman Naseem, and Mehwish Nasim. 2023b. Nehate: Large-scale annotated data shedding light on hate speech in nepali local election discourse. In *ECAI 2023*, pages 2346–2353. IOS Press.

Durgesh Verma and Abhinav Kumar. 2025. Multimodal Deep Learning for Detection of Hate, Humor, and Stance in Social Discourse on Marginalized Communities. In *Proceedings of The Eighth Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

A Related Work

As memes become a popular way to express opinions on social and political issues, researchers are paying more attention to analyzing both their text and images to detect hate, humor, and ideologically charged content. The Hateful Memes Challenge (Kiela et al., 2020) introduced one of the earliest benchmark datasets containing synthetic memes with contrastive image-text signals targeting protected categories such as race, religion, and gender. Subsequent datasets like Harm-C and Harm-P (Pramanick et al., 2021a,b) captured real-world political and COVID-related memes, annotated across varying degrees of harmfulness and target categories. CrisisHateMM (Bhandari et al., 2023) similarly focused on memes related to the Russia-Ukraine conflict and included hate speech target classification. Beyond hate detection, other efforts have targeted different aspects of meme communication: Suryawanshi et al. (2020) introduced MultiOFF for offensive meme detection using data from Reddit and Instagram; Tanaka et al. (2022) proposed a humor detection dataset by extracting memes without interpersonal bias; and DisinfoMeme (Qu et al., 2022) focused on disinformation, annotating memes from movements like BLM and Veganism. Parallel efforts have explored multi-aspect datasets capturing a wider spectrum of linguistic phenomena, Gautam et al. (2020) annotated the MeToo movement-related tweets across dialogue acts, sarcasm, stance, and hate; Dacon et al. (2022) labeled LGBTQ-related Reddit comments for toxicity and identity attacks; and Ousidhoum et al. (2019) provided a multilingual corpus annotated for hate, offensiveness, stance, and sentiment. Recent shared tasks have explored these challenges further: the CASE 2024 Climate Activism task (Thapa et al., 2024a) annotated tweets for stance, hate speech, and humor; the NAET dataset (Rauniyar et al., 2023) collected Nepali anti-establishment tweets with multi-aspect labels including satire, hate, and hope speech; and NEHATE (Thapa et al., 2023b) focused on identifying hate speech and its targets in Nepali election discourse. Similarly, the GameTox dataset (Naseem et al., 2025) introduced token-level and intent-level annotations for toxicity in gaming chats. While most prior work focuses on single aspects or monolingual textual analysis, our task offers a multimodal and multi-aspect benchmark covering hate, stance, humor, and targeted hate enabling richer exploration of social discourse through memes.

B Evaluation and Competition

This section outlines the overall framework of our shared task, the evaluation methodology, competition structure, and key logistical information.

B.1 Evaluation Metrics

We employed a suite of standard classification metrics to evaluate performance: accuracy, precision, recall, and the macro F1-score. The official ranking of the participating teams on the final leaderboard was determined based on their macro F1-score.

B.2 Competititon Setup

The shared task was hosted on the CodaLab platform[1], which provided a standardized environment for all participants. The competition was structured into two primary phases: a development phase and a final testing phase, followed by a peer-reviewed paper submission process.

Registration. A total of 89 participants registered for the shared task, which shows strong interest from individuals across diverse professional backgrounds. Geographic diversity was also notable, as indicated by the wide range of email domain affiliations. Of the registered participants, 21 teams submitted their final prediction outputs.

Competition Timelines. The competition officially commenced on April 8, 2025, with the release of the training and evaluation datasets. This initial phase allowed participants a full month to familiarize themselves with the data, develop their models, and perform internal validation.

The final testing phase began on May 8, 2025, with the release of the test set, for which the ground truth labels were withheld. Participants had over two months to apply their systems to the test data, with the testing period concluding on July 12, 2025.

Following the completion of the testing phase, teams were invited to document their methodologies in a system description paper, with submissions due by Aug 4, 2025. These papers underwent a formal review process, and notifications were sent to authors on August 17, 2025. Authors of conditionally accepted papers were given until August 24, 2025, to submit their revised versions, with final notifications sent on August 25, 2025. The camera-ready versions of all accepted papers were due on August 30, 2025. The shared task will culminate with presentations at the CASE Workshop from September 11-13, 2025. This structured timeline, coupled with continuous support for any technical issues, was designed to facilitate a productive and engaging research environment for all participants.

[1] https://codalab.lisn.upsaclay.fr/competitions/22463