Overfitters@CASE2025: Multimodal Hate Speech Analysis Using BERT and RESNET

Bidhan Chandra Bhattarai[1*], Ishan Maharjan[1*], Dipshan Pokhrel[1], Rabin Thapa[1]
[1]IIMS College, Kathmandu, Nepal
{bidhan, ishan, dipshan, rabin}@iimscollege.edu.np

*These authors contributed equally to this work

Abstract

Marginalized socio-political movements have become focal points of online discourse, polarizing public opinion and attracting attention through controversial or humorous content. Memes, play a powerful role in shaping this discourse both as tools of empowerment, and as vessels for ridicule or hate. The ambiguous and highly contextual nature of these memes presents a unique challenge for computational systems. In this work we try to identify these trends. Our approach leverages the BERT+ResNet(BERTRES) model to classify the multimodal content into different categories based on different tasks for the Shared Task on Multimodal Detection of Hate Speech, Humor, and Stance in Marginalized Socio-Political Movement Discourse at CASE 2025. The task is divided into four sub-tasks: subtask A focuses on detection of hate speech, subtask B focuses on classifying the targets of hate speech, subtask C focuses on classification of topical stance and subtask D focuses on detection of intended humor. Our approach obtained a 0.73 F1 score in subtask A, 0.56 F1 score in subtask B, 0.6 F1 score in subtask C, 0.65 F1 score in subtask D.

1 Introduction

In the constantly evolving digital landscape, social media have managed to become a constant and integral means of information exchange and communication. While social media platforms have certainly facilitated an increase in online engagement, they have unsurprisingly been a hotbed for online abuse, cyberbullying, and the proliferation of hate speech.

Hate speech refers to any expression, whether spoken, written or nonverbal, that targets, insults or dehumanizes individuals based on aspects of their social identity, including race, religion, ethnicity, gender or sexual orientation. It encompasses forms of communication likely to incite, justify or reinforce harm such as violence, discrimination or systemic oppression particularly when these expression arise from entrenched power dynamics and historical injustices. (Ruscher, 2025). Hence, the spread of hate speech across the internet through the means of social media platforms has become a complex societal issue (Shiwakoti et al., 2024; Thapa et al., 2023b; Jafri et al., 2024, 2023). Memes, which integrate images with accompanying text in the form of captions, have been utilized to propagate hate speech. Detecting hate speech in multimodal content, such as memes requires more than just text or image analysis. Unimodal algorithms that focus on either image or text analysis fail at understanding contexts and nuances, making them ineffective when analyzing multimodal content. As a result, detecting hate speech in such multimodal content requires more than traditional text or image analysis.

The Hateful Memes Challenge (Kiela et al., 2020) was one of the significant attempts to benchmark systems capable of tackling this complex issue. Since then, tasks such as those organized in the CASE workshop series (Thapa et al., 2023a, 2024c) have continued to improve the field by incorporating more fine-grained tasks. These include stance detection, humor recognition, and classification of hate speech targets.

The CASE 2025 Shared Task (Thapa et al., 2025) uses **PrideMM** (Shah et al., 2024), which is a dataset of memes focused on content related to LGBTQ+ rights and marginalized sexual minorities. This dataset was introduced in the MemeCLIP paper. This paper includes annotated training, validation, and test sets, making it a comprehensive resource to understand memes through the lens of hate, humor, stance, and target.

Our submission to the CASE 2025 shared task introduces **BERTRES**. It is a multimodal fusion model that combines textual features from Bidirectional Encoder Representations from Transform-

Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025, pages 76–82, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-009

ers (BERT) (Devlin et al., 2019) and visual embeddings from ResNet-50 (He et al., 2015). Our BERTRES model performed strongly on Subtasks A - hate speech and D - humor, with test scores of 0.7377 and 0.6533, respectively, but struggled on Subtasks B - target and C - stance due to label imbalance and semantic accuracy.

2 Related Work

There has been substantial research over the years on various aspects of hate and toxicity detection (Thapa et al., 2024b; Naseem et al., 2025; Rauniyar et al., 2023). More recently, multimodal hate speech detection has gained serious traction, due to the popularity of memes as a means of communication and information exchange (Thapa et al., 2024a). Social media platforms have become battlegrounds, where memes are used not just for laughs but to push political agendas, express support or spread hate, often simultaneously.

One of the major contributions in this area is the Hateful Memes Challenge at NeurIPS 2020 (Kiela et al., 2020), which introduced a dataset consisting of 10,000 meme examples. This dataset helped create an environment to create state of the art models in this field.

Newer datasets, such as MemeCLIP (Shah et al., 2024), leveraged vision language pretraining via CLIP to improve meme classification in low-context environment. It provided the PrideMM dataset, the primary dataset used in CASE 2025. Prior datasets like the CrisisHateMM dataset (Bhandari et al., 2023) emphasized the importance of distinguishing between directed and undirected hate.

Furthermore, multimodal approaches have matured significantly in recent times. MemeFier (Koutlis et al., 2023) used transformer based fusion with task specific modules to enhance interpretability. Hate-CLIPper (Kumar et al., 2022) employed attention based alignment for robustness in zero-shot settings. Aggarwal et al. (2024) emphasized that in generalization scenarios, textual features mainly dominate model performance. HateSieve (Su et al., 2024) introduced cross-modal contrastive objectives for joint detection and segmentation of hateful elements.

Recognizing this shift, the CASE workshop series has been pivotal in multimodal meme analysis. Earlier shared tasks like CASE 2023 Recognizing this shift, the CASE workshop series has been leading the charge. Early shared tasks like CASE 2023 (Thapa et al., 2023a) and CASE 2024 (Thapa et al., 2024c) focused on crisis events namely the Russia-Ukraine conflict. These tasks highlighted how hate speech evolves during geopolitical turmoil and showed the need for contextually aware multimodal models which can analyze both image and textual content simultaneously.

The CASE 2025 shared task (Thapa et al., 2025) introduces a unified evaluation benchmark using the PrideMM dataset. It brings together a diverse set of challenges such as hate speech, stance and humor together under a single shared task. Meanwhile, broader insights into socio-political event detection, including multimodal and cross-lingual trends are documented in the workshop overview paper (Hürriyetoğlu et al., 2025).

3 Dataset and Task

The **PrideMM** dataset is the primary dataset provided for the shared task CASE 2025. It comprise of a curated collection of memes annotated for four distinct subtasks namely, hate speech detection, target classification, stance detection and humor detection. The dataset is a static image with an accompanying or an overlaid caption, which reflects the multimodal nature of memes.

The shared task is structured into the following four subtasks:

- **Subtask A: Hate Speech Detection** — Identify whether the meme contains hateful content, distinguishing between *hate* and *non-hate* expressions.

- **Subtask B: Target Classification** — Determine the entity or group targeted by the hate speech, categorized as *Undirected*, *Individual*, *Community*, or *Organization*.

- **Subtask C: Stance Classification** — Assess the stance conveyed towards the identified target, classified as *Support*, *Neutral*, or *Oppose*.

- **Subtask D: Humor Detection** — Classify whether the meme is intended to be binary classification of humor, an important dimension given humor's complex role in meme communication.

Analyzing the class distribution reveals mild imbalances, mainly in Subtasks B and Subtasks C, where a large majority of samples are labelled as neutral or non-targeted. Humor detection subtask, on the other hand, benefits from intentional oversampling and a more balanced representation is reached, enabling fairer model evaluation.

This shared task is motivated by the challenges inherent in real-world content moderation, where the interplay of multimodal cues, subtle biases, and the socio-political context complicate automated analysis. The PrideMM dataset and the associated task design offer a novel and rigorous benchmark for evaluating the robustness, fairness, and interpretability of systems aimed at socio-political meme understanding and moderation.

Subtask	Label	Split	Samples
A: Hate Speech	No Hate	Train	2065
	Hate	Train	1985
	-	Val	506
	-	Test	507
B: Target	Undirected	Train	617
	Organization	Train	238
	Individual	Train	199
	Community	Train	931
	-	Val	248
	-	Test	249
C: Stance	Support	Train	1527
	Oppose	Train	1357
	Neutral	Train	1166
	-	Val	506
	-	Test	507
D: Humor	Humor	Train	2737
	No Humor	Train	1313
	-	Val	1012
	-	Test	507

Table 1: Dataset Sample Distribution Across Subtasks

4 Methodology

We present **BERTRES**, a multimodal fusion architecture designed to capture and integrate the rich semantic cues inherent in both textual and visual component of memes. We understand that memes combine image and text in such a way that they jointly convey complex messages, and just analyzing memes using uni-modal analysis is not sufficient to understand the context. As a result, our approach leverages specialized encoders for each modality before fusing their representations for multimodal classification.

The textual content of each meme is tokenized and passed through a pre-trained BERT base model (Devlin et al., 2019). BERT helps us extract the [CLS] token embedding, which is a 768-dimensional vector that encapsulates the overall semantic meaning of the caption. This embedding can be regarded as a summary of the textual information, and it is generally nuanced and context-dependent.

At the same time, the visual information undergoes preprocessing to adhere to the input requirements of a ResNet-50 Convolutional Neural Network (He et al., 2015). Resnet-50 is pre-trained on large scale image recognition tasks. The output of the ResNet-50 is a 2048-dimensional feature vector, which captures visual patterns and contextual details within the meme to contribute to its meaning.

BERTRES works by fusing these two modality specific embeddings. It concatenates the text embedding with the visual embedding to obtain a comprehensive feature vector, that represents the multimodal content of the meme. The vector is then passed through a series of fully connected layers incorporating ReLU activations, batch normalization and dropout regularization that help the model learn complex interactions between modalities while mitigating overfitting.

The diversity of classification tasks inherent to meme analysis addressed by the model employing separate classification heads for each of the four subtasks. This helps the model to share feature extraction layers to learn generalized multimodal representations, while at the same time enables each head to specialize in its respective classification objective.

Training uses the Adam optimizer with a finely tuned learning rate of 2×10^{-5}. To overcome class imbalance, especially seen in target and stance classification subtasks, we use class-weighted cross entropy loss to ensure the model fairly attends to underrepresented classes. This method, along with dropout and batch normalization, promotes robustness and improves generalization across subtasks.

4.1 BERTRES Model Architecture

The model architecture is consistent across subtasks, adapting only the final classification layer to the number of classes for each task (2 for hate speech and humor detection, 4 for target classification, and 3 for stance detection). The key components include:

- **Image Processing:** The visual encoder is a pre-trained ResNet-50 network. We remove

Table 2: Training Setup for Each Subtask

Subtask	Model	BS	Ep.	LR
A: Hate Speech	ResNet+BERT	16	2	1×10^{-5}
B: Target	ResNet+BERT	16	6	2×10^{-5}
C: Stance	ResNet+BERT	16	6	2×10^{-5}
D: Humor	ResNet+BERT	16	2	1×10^{-5}

its original classification layer and replace it with an identity mapping to extract a 2048-dimensional feature vector. This vector is then linearly projected down to a 768-dimensional embedding to align with the textual feature space, facilitating effective fusion.

- **Text Processing:** Textual input is encoded using a pre-trained BERT base uncased model. We utilize the [CLS] token embedding from the last hidden state as a compact yet rich representation of the entire caption.

- **Feature Fusion and Classification:** The concatenated multimodal feature vector comprising of 1536-dimensional after projection passes through a dropout layer with dropout rate 0.5 before entering a two-layer fully connected classifier. The first layer reduces dimensionality from 1536 to 512 with ReLU activation, followed by the output layer which maps to the appropriate number of classes.

While fine-tuning both BERT and ResNet-50 does allow the model to adapt the representations to the specific nuances of meme data, there is always a risk of overfitting considering the small size and imbalanced nature of the dataset. We try to mitigate this through dropout, batch normalization, and class-weighted losses, balancing adaptability with generalization.

4.2 Training Setup

Table 2 details the training configurations employed for each subtask. Consistent batch sizes and carefully chosen epochs reflect the balance between training efficiency and performance, while the learning rates and optimization strategies are tuned to ensure convergence without overfitting.

5 Results & Discussion

Table 3 presents BERTRES's final leaderboard performance across the four subtasks in the CASE 2025 shared task. Each subtask posed unique challenges, ranging from implicit hate expression to ambiguous humor which required robust multimodal analysis to achieve reasonably accurate predictions.

Table 3: Final Leaderboard Performance of BERTRES

Subtask	F1 Score	Rank
A: Hate Speech Detection	0.7377	15
B: Target Classification	0.5628	6
C: Stance Classification	0.6015	9
D: Humor Detection	0.6533	14

Subtask A required the identification of hate speech in memes. BERTRES achieved an F1 score of 0.7377 and ranked 15th. This relatively lower performance can be attributed to the subtle nature of implied hate and sarcasm, a task which is inherently difficult to model without contextual meta-data.

Subtask B required target identification and BERTRES, securing 6th position with an F1 score of 0.5628. We attribute this result to our use of class-weighted loss and balanced representation learning, which helped mitigate the skewed label distribution among target types. There is a need for more robust algorithms to improve the prediction in this task.

Subtask C focused on identifying stance (support, oppose, neutral), BERTRES ranked 9th, obtaining an F1 of 0.6015. The results suggest that while BERTRES captured some of the underlying intent in meme discourse, it struggled with cases involving satire or ambiguous sentiment.

Subtask D required humor detection proved to be particularly challenging. Our system scored an F1 of 0.6533, placing 14th. Humor's subjective and culturally grounded nature, coupled with limited contextual cues, made it difficult for the model to generalize.

Overall, BERTRES demonstrated consistent mid-tier performance, with its strongest results in target classification and respectable scores in the remaining subtasks. These results highlight both the strengths of a fusion-based architecture and the inherent complexities of multimodal socio-political content moderation.

6 Conclusion

This paper presented our approach to the CASE 2025 Shared Task on Multimodal Detection of Hate Speech, Humor and Stance in Marginalized

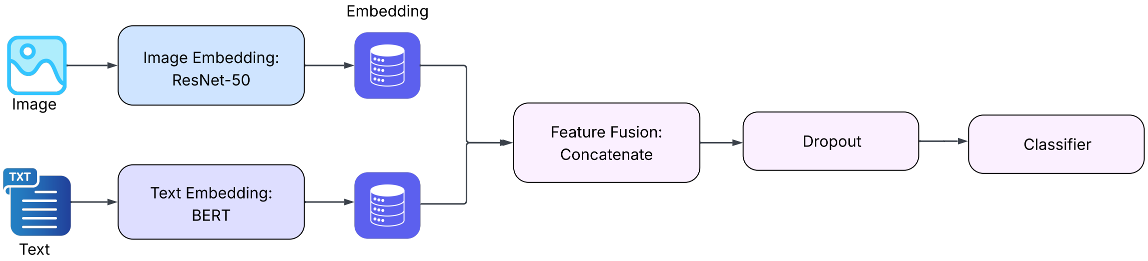

Figure 1: System overview of the BERTRES model combining textual and visual modalities for meme classification across four subtasks.

Socio-Political Movement Discourse. We came up with **BERTRES**, a lightweight but an effective multimodal model that fuses representations from BERT and Resnet-50 model. Our system performed strongly, particularly in hate speech and target classification, demonstrating competitive generalization capabilities across diverse meme types.

Even though the performance in humor and stance detection were modest, we identified important challenges such as semantic ambiguity and cultural disparity that hindered classification accuracy. Our analysis highlights the need for more context-aware modeling and improved representation of nuanced information and sentiment in multimodal content.

Future work will focus upon the improvement of fusion strategy, by incorporating contrastive learning techniques, and adapting prompt-based methods that can dynamically interpret memes within their socio-political context.

7 Limitations

BERTRES demonstrated overfitting on subtasks with imbalanced or sparse label distributions, especially in the case of stance classification.Memes heavily depend upon external cultural or political context not present in the text or image alone. Additionally, though our model is pretrained it lacks access to real-world information, which made it difficult to understand the context. Furthermore, the current fusion mechanism concatenated image and text embeddings without inter-modal attention, which limited adaptability in ambiguous or sarcastic and humorous memes.

8 Ethical Considerations

While working with data related to hate speech and marginalized communities, we came across some important ethical concerns. Firstly, PrideMM includes real-world memes that reflect hate, discrimination, and political rhetoric. Researchers should handle such data respectfully and avoid causing harm. Also model predictions can reflect annotation and training biases, mainly in underrepresented subgroups. Careful evaluation and auditing is essential before deployment in real-world content moderation systems.Even though these models are developed for research, these models can be misapplied for surveillance and censorship. Transparency, reproducibility and appropriate safeguards are required to combat the potential for misuse. The dataset pertains to LGBTQ+ issues, any future extensions or applications should involve stakeholders from those groups so that fairness and transparency can be ensured.

We aim to support positive use cases such as harmful content detection and inclusive moderation tools, but future research should continue to foreground ethical awareness alongside technical progress to ensure that the ethical standards are always met.

References

Shreya Aggarwal, Jasdeep Singh, Harshit Chauhan, Aditya Mittal, and Mohit Bansal. 2024. Text vs. vision-language models for generalizable multimodal hate speech detection. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and

Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

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 Sociopolitical Events from Text (CASE 2025)*.

Farhan Ahmad Jafri, Kritesh Rauniyar, Surendrabikram Thapa, Mohammad Aman Siddiqui, Matloob Khushi, and Usman Naseem. 2024. Chunav: Analyzing hindi hate speech and targeted groups in indian election discourse. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Farhan Ahmad Jafri, Mohammad Aman Siddiqui, Surendrabikram Thapa, Kritesh Rauniyar, Usman Naseem, and Imran Razzak. 2023. Uncovering political hate speech during indian election campaign: A new low-resource dataset and baselines. *arXiv preprint arXiv:2306.14764*.

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.

Marios Koutlis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2023. Memefier: A two-stage fusion framework for multimodal meme classification. In *Proceedings of the 2023 International Conference on Multimedia Retrieval (ICMR)*, pages 210–218. ACM.

Abhinav Kumar, Aishwarya Jaiswal, Pavan Kapanipathi, and Gerald Tesauro. 2022. Hate-clipper: Multimodal hate speech detection using cross-modal interaction matrices. *arXiv preprint arXiv:2210.12357*.

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.

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.

Janet B. Ruscher. 2025. *Hate Speech*. Elements in Applied Social Psychology. Cambridge University Press.

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

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).

Zihan Su, Yong-Hwi Lee, Xiang Zhang, and Jiebo Luo. 2024. Hatesieve: Segmenting and detecting hateful content in multimodal memes via contrastive learning. *arXiv preprint arXiv:2402.08033*.

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, Farhan Ahmad Jafri, Kritesh Rauniyar, Mehwish Nasim, and Usman Naseem. 2024a. Ruhate-mm: Identification of hate speech and targets using multimodal data from russia-ukraine crisis. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1854–1863.

Surendrabikram Thapa, Kritesh Rauniyar, Farhan Ahmad Jafri, Shuvam Shiwakoti, Hariram Veeramani, Raghav Jain, Guneet Singh Kohli, Ali Hürriyetoğlu, and Usman Naseem. 2024b. 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. 2024c. 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.

Surendrabikram Thapa, Siddhant Bikram Shah, Kritesh Rauniyar, Shuvam Shiwakoti, Surabhi Adhikari, Hariram Veeramani, Kristina T. Johnson, Ali Hürriyetoğlu, Hristo Tanev, and Usman Naseem. 2025. 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).*