MMFusion@CASE 2025: Attention-Based Multimodal Learning for Text-Image Content Analysis

Prerana Rane
IEEE Senior Member / California, USA
preranarane@ieee.org

Abstract

Text-embedded images, such as memes, are now increasingly common in social media discourse. These images combine visual and textual elements to convey complex attitudes and emotions. Deciphering the intent of these images is challenging due to their multimodal and context-dependent nature. This paper presents our approach to the Shared Task on Multimodal Hate, Humor, and Stance Detection in Marginalized Movement at CASE 2025[1]. The shared task focuses on four key aspects of multimodal content analysis for text-embedded images: hate speech detection, target identification, stance classification, and humor recognition. We propose a multimodal learning framework that uses both textual and visual representations, along with cross-modal attention mechanisms, to classify content across all tasks effectively.

1 Introduction

The prevalent use of text-embedded images, particularly memes, in social media has raised new challenges in detecting harmful content. Traditional text-only methods are not effective in capturing semantic context when images and text work together to convey complex negative messages. Multimodal approaches perform better than unimodal methods in detecting harmful content, which often relies on the interaction between visual and textual elements (Kiela et al., 2020).

Previous editions of the multimodal hate speech event detection shared tasks (Thapa et al., 2024, 2023) have addressed challenges in detecting hate speech in text-embedded images related to sociopolitical events. The Shared Task on Multimodal Hate, Humor, and Stance Detection in Marginalized Movement at CASE 2025 (Thapa et al., 2025; Hürriyetoğlu et al., 2025) introduces multimodal classification with four subtasks, each targeting a different aspect of online discourse: (A) detection of hate speech, (B) classification of hate speech targets, (C) stance classification toward marginalized movements, and (D) humor recognition. This paper presents our system, which uses a multimodal architecture combining text and image encoders with cross-modal attention mechanisms to extract relevant features.

2 Related Work

The detection of harmful or sensitive content in text-embedded images has gained attention with the rise of social media. Recent work highlights the challenges in automating hate speech detection due to complex linguistic cues and implicit expressions of hate. (Parihar et al., 2021). Early work on hate speech detection focused on textual data (Davidson et al., 2017; Waseem and Hovy, 2016). However, text-embedded images require a multimodal analysis of both textual and visual cues to understand implicit meanings and cultural references common in social media discourse (Kiela et al., 2020).

Prior research in stance classification focused on deciphering explicit stance indicators in text (Mohammad et al., 2016). More recent work leverages transformer models for text (Küçük and Can, 2020) to capture contextual nuances in stance detection. Humor recognition requires an understanding of context, cultural nuances, and figurative language (Annamoradnejad and Zoghi, 2020). Recent work has explored the use of contextual embeddings and attention mechanisms to capture the subtle linguistic patterns that characterize humorous content (Weller and Seppi, 2020)

Visual deciphering of harmful content has used convolutional neural networks such as ResNet (He et al., 2016) to extract features from images.

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

Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025, pages 115–122, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-014

Transformer-based models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have been used to extract contextual embeddings from text. Cross-modal attention captures fine-grained interactions between different modalities, such as text and images (Chen et al., 2020; Li et al., 2019). These attention-based fusion mechanisms are essential for recognizing subtle forms of harmful content, sarcasm, or humor.

Misclassifications or errors in sensitive content can lead to serious consequences, which needs robustness in multimodal classification systems (Larson, 2017). Test-time augmentation (TTA) has shown promise in computer vision (Wang et al., 2019) but its application in multimodal applications have been limited.

Our work addresses the CASE 2025 Shared Task by proposing a multimodal architecture designed for hate speech detection, target identification, stance classification, and humor recognition. The multimodal system integrates transformer-based text encoders (BERT and RoBERTa) with CNN-based image encoders (ResNet variants). It uses a cross-modal attention fusion mechanism to capture fine-grained interactions between text and image features. We incorporate TTA to enhance prediction stability and reduce errors on unseen data across all tasks.

3 Dataset & Task Description

We have used the PrideMM dataset (Shah et al., 2024). PrideMM is a dataset containing 5,063 text-embedded images related to the LGBTQ+ movement collected from Facebook, Twitter, and Reddit. The annotation scheme was adopted from (Bhandari et al., 2023). Table 1 presents the dataset size for training, validation and testing for each task.

Subtask	Train	Val	Test
A	4050	506	507
B	1985	248	249
C	4050	506	507
D	4050	506	507

Table 1: PrideMM dataset sizes for each task

Data pre-processing included text cleaning (e.g., URL removal, normalization of whitespace and punctuation, and conversion of hashtags and mentions) and image normalization using ImageNet statistics.

3.1 Subtask A: Hate Speech Detection

Hate Speech Detection involves binary classification to determine the presence of hate speech in text-embedded images. Given an image paired with a textual description, the task requires the system to classify the content as "No Hate" or "Hate". The training dataset for subtask A has a nearly balanced class distribution with 51.0% No Hate (2065 images) and 49.0% Hate (1985 images).

3.2 Subtask B: Target Identification

Target Identification involves classifying the targets in text-embedded hate speech content. Given an image that has already been identified as containing hate speech, the task requires the system to classify the content into one of four target categories: "Undirected," "Individual," "Community," or "Organization." Undirected hate speech contains hateful content without targeting specific entities. The Individual, Community, and Organization categories require the system to distinguish between personal attacks, group-targeted hate, and institutional criticism, respectively. The training dataset for subtask B contains 31.1% Undirected (617 images), 10.0% Individual (199 images), 46.9% Community (931 images) and 12.0% Organization (238 images).

3.3 Subtask C: Stance Classification

Stance Classification involves classifying stance in text-embedded images. The task requires the system to classify the content into three stance categories: "Neutral", "Support" and "Oppose". The training dataset for subtask C contains 28.8% Neutral (1166 images), 37.7% Support (1527 images), and 33.5% Oppose (1357 images).

3.4 Subtask D: Humor Recognition

Humor Recognition involves binary classification of text-embedded images to determine if the content contains humor, sarcasm, or satire. The task requires the system to classify the content as "No Humor" or "Humor". The training dataset for subtask D contains 32.4% No Humor (1313 images) and 67.6% Humor (2737 images).

4 Methodology

For all tasks, our multimodal architecture consists of three main components: (1) text encoder (2) image encoder and (3) cross-modal or self-attention mechanism. We have used BERT, RoBERTa and

DialoGPT to extract text features, and ResNet to extract image features. BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) are transformer-based models designed to capture deep contextual dependencies in text. DialoGPT (Zhang et al., 2020) is a variant of GPT-2 fine-tuned on large-scale dialogue datasets to better model conversational language. ResNet (He et al., 2016) is a deep convolutional neural network that introduces residual connections to ease the training of very deep models.

For Hate Speech Detection and Target Identification, we used the RoBERTa-base model for the text, with a maximum sequence length of 256 tokens and the CLS token embeddings (768 dimensions) as the primary feature representation. For the images, we used a ResNet50 model pre-trained on ImageNet, removing the final classification layer and extracting a 2048-dimensional feature vector from the global average pooling layer. Both text and image features were projected into a 512-dimensional space using linear transformations and then combined using an 8-head multi-head attention mechanism. The fused features were passed through a multilayer perceptron (MLP) classifier. The output of Hate Speech Detection is a binary classification of No Hate (0) or Hate (1). The output of Target Identification is Undirected (0), Individual (1), Community (2), or Organization (3). The high-level system design for Hate Speech Detection and Target Identification is shown in Figure 1.

For Stance Classification, we used an ensemble of multimodal classifiers to combine textual and visual features. Each model in the ensemble processes text and image modalities through separate branches before fusing the features via a shared projection layer. For text, we use RoBERTa-base and BERT-base-uncased as our encoders, extracting CLS token embeddings with a maximum sequence length of 128 tokens. These embeddings are linearly projected to a 256-dimensional space for cross-modal fusion. For images, we use ResNet18 and ResNet34 pretrained on ImageNet, from which we extract global average pooled convolutional features.These visual representations are projected into the same 256-dimensional feature space. We use a simple attention mechanism to learn dynamic weighting between text and image features. The fused representation is created by concatenating the projected text and image features, followed by classification through a fully connected layer. Final predictions are generated through probability averaging. The output is Neutral (0), Support (1) or Oppose (3). The high-level system design for Stance Classification is shown in Figure 2.

For Humor Recognition, the text is processed using DialoGPT-medium, chosen for its ability to handle conversational and informal language in social media humor. Tokenized sequences are truncated or padded to a maximum of 196 tokens. From the encoder, we extract token embeddings, apply mean pooling over the sequence length, and project the resulting representation into a 512-dimensional feature vector. For images, we use the ResNet50 model. The extracted 2048-dimensional features are projected to a 512-dimensional space for cross-modal fusion. We applied self-attention mechanisms independently on text and image features. We then used cross-modal attention, where text features act as the query and image features as the key-value pairs. A gating mechanism adaptively weights the text and image features. The final fused representation, formed by combining gated text, gated image, and cross-modal attention outputs (3 $\times$ 512 dimensions), is passed through a multi-layer classifier with progressively reduced dimensions. The output is a binary classification of No Humor (0) or Humor (1). The high-level system design for Humor Recognition is shown in Figure 3.

The choice of architectures for the subtasks was guided by task-specific requirements and empirical performance. Subtasks A and B use RoBERTa and ResNet50 for binary and multi-class classification. Subtask C employs an ensemble strategy to address the severe class imbalance in stance detection. For Subtask D, DialoGPT replaced RoBERTa to capture conversational patterns and humor cues.

5 Results & Discussion

All experiments were conducted using the Hugging Face Transformers library for access to RoBERTa-base, BERT-base, and DialoGPT-medium models. The multimodal architectures were implemented in PyTorch 1.13 with NVIDIA CUDA support. F1 score is the primary evaluation metric for all the tasks.

5.1 Experiment Setup

For hate speech detection and target identification, we used focal loss (γ = 2.0) to focus on hard-to-classify samples, using AdamW optimizer (learning rate of 1e-5, weight decay of 0.01) and a linear

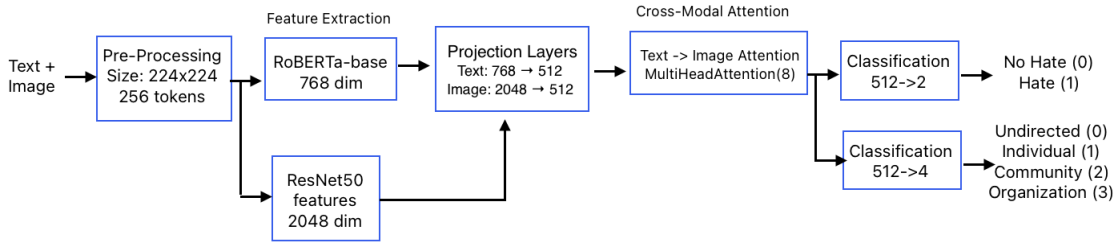

Figure 1: High-level system design for Hate Speech Detection and Target Identification. Text and image inputs are processed through separate encoders (RoBERTa for text and and ResNet50 for images), followed by a cross-modal fusion layer and classification.

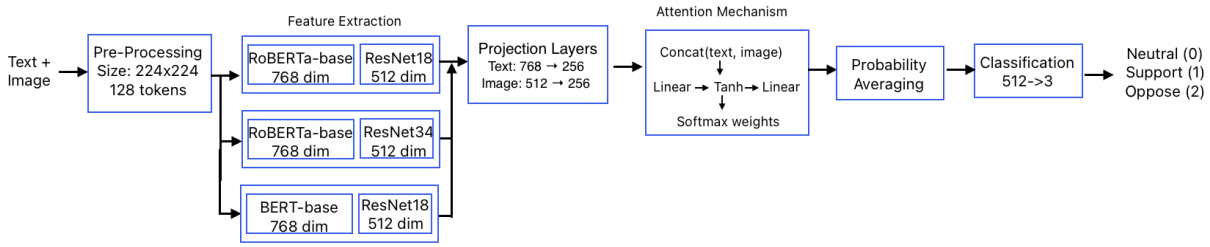

Figure 2: High-level system design for Stance Classification. Three base models (RoBERTa-ResNet18, RoBERTa-ResNet34, and BERT-ResNet18) process multimodal features. Their output probability distributions are averaged to produce the final stance prediction across the three classes: Neutral, Support, and Oppose.

warmup schedule followed by a linear decay. For target identification, we used focal loss with class-specific weighting to handle residual imbalance. We trained the model for eight epochs with a batch size of 12, applying gradient clipping (norm $\leq$ 1.0) to stabilize updates. We also used a test-time augmentation (TTA) strategy that generated five variants of each test image (original, horizontal flip, brightness/contrast, rotation, and color adjustment). The softmax probabilities across all augmentations were averaged before making a final prediction to enhance classification.

For stance classification, we train three models. The first model uses RoBERTa-base with ResNet18, the second model combines RoBERTa-base with ResNet34, and the third model uses BERT-base with ResNet18. These models are trained independently with different random seeds (42, 123, and 456) to encourage diversity within the ensemble. We use a label-smoothing, class-weighted cross-entropy loss to address the moderate class imbalance in the dataset. The class weights are computed inversely proportional to class frequencies and applied during optimization. All models are trained using the AdamW optimizer with a learning rate of 2e-5, weight decay of 0.01, and gradient clipping at a maximum norm of 1.0 for six epochs. To reduce overfitting, dropout is applied in the fusion layers (0.3) and the classifier (0.15), while the embedding layers are partially frozen during the initial training phases for stability. We perform ensemble prediction by averaging the probability outputs of the three trained models and selecting the class with the highest probability.

For humor recognition, we used focal loss (α=1, γ=2), which reduces the effect of class imbalance. Optimization is performed with AdamW (learning rate = 1e-5, weight decay = 0.01) and a cosine annealing schedule for 15 epochs. We used a batch size of 12 and gradient clipping (maximum norm = 1.0) for stability. Regularization strategies include dropout (0.3 across layers), partial freezing of DialoGPT embedding layers, and test-time augmentation as described previously.

5.2 Results

Table 2 presents the evaluation results for all the tasks on the test dataset.

Task	Recall	Precision	F1	Accuracy
A	0.779	0.781	0.778	0.779
B	0.550	0.565	0.553	0.590
C	0.611	0.612	0.608	0.611
D	0.648	0.700	0.658	0.733

Table 2: Evaluation results for tasks

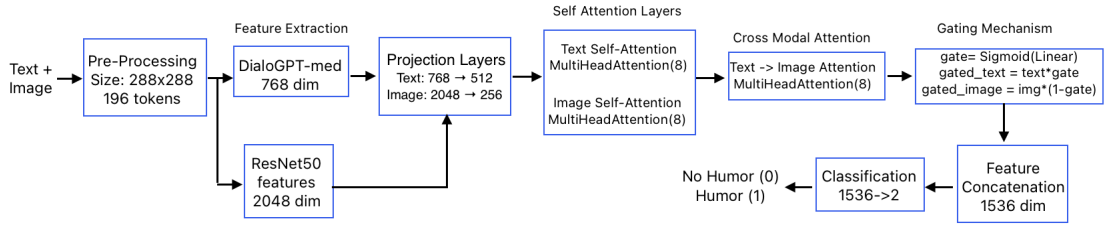

Figure 3: High-level system design for Humor Recognition. DialoGPT-medium for text encoding and ResNet50 for image encoding, followed by cross-modal attention layers fused through a gating mechanism and fed into a multi-layer classifier for binary humor prediction.

The hate speech detection task achieved an F1 score of 0.778, precision (0.781), recall (0.779), and accuracy (0.779), suggesting that the multimodal architecture effectively captured the visual and textual patterns. The target identification task achieved an F1 score of 0.553, precision (0.565) and recall (0.550). The lower F1 score may indicate that the model struggles with certain class boundaries. The precision-recall gap of 0.015 suggests conservative predictions. While our performance falls short of recent shared task winner (Wang and Markov, 2024) (CLTL: 87.27% and 80.05% respectively) and the CLIP baseline (78.60% and 61.50%), our results demonstrate competitive performance within the challenging multimodal classification domain. The performance gap highlights the difficulty of these tasks and suggests directions for future improvement in fusion mechanisms and using pretraining strategies used by top-performing systems.

The stance classification task achieved an F1 score of 0.608, with precision (0.612) and recall (0.611), showing consistent performance across all three stance categories (Neutral, Support, Oppose). The low difference between precision and recall suggests that our approach balanced the moderately imbalanced class distribution (28.8% Neutral, 37.7% Support, 33.5% Oppose).

The humor detection task yielded an F1 score of 0.658 with higher precision (0.700) than recall (0.648), indicating that our model is conservative in predicting humor, preferring to avoid false positives. The accuracy of 0.733 reflects higher classification performance, while the precision-recall gap suggests that the focal loss strategy and cross-modal attention mechanisms successfully addressed the class imbalance (67.6% humor vs 32.4% no humor) by being more selective in humor predictions.

To validate our task-specific architecture choices, we compared multiple approaches across subtasks. For hate speech detection, RoBERTa+ResNet50 with cross-modal attention achieved the best performance (F1=0.778), outperforming ensemble methods (F1=0.726). Target identification showed similar patterns with RoBERTa+ResNet50 (F1=0.553) exceeding ensemble approaches (F1=0.547). For stance classification, systematic comparison showed that individual models struggled: RoBERTa+ResNet50 (F1=0.559), DialoGPT+ResNet50 (F1=0.533), and BERT-base+ResNet50 (F1=0.443). This performance degradation led to adopting an ensemble approach with simple attention, achieving F1=0.608. For humor detection DialoGPT+ResNet50 (F1=0.658) outperformed both RoBERTa+ResNet50 (F1=0.646) and ensemble methods (F1=0.630).

5.3 Error Analysis

Figures 4-7 show the error patterns across the subtasks, based on the varying complexity of each classification challenge. Subtask A (Hate Speech Detection) achieved 188/258 (72.9%) correct ”No Hate” predictions and 206/249 (82.7%) correct ”Hate” predictions. The primary error pattern shows 70 false positives, where non-hateful content was misclassified as hateful, suggesting that the model may be sensitive to certain linguistic patterns or visual elements associated with hate speech. For example, ”gay marriage shouldn’t exist, it should just be considered marriage” has been incorrectly classified as Hate.

The model for Subtask B (Target Classification) struggles with distinctions between target categories. The ”Individual” class shows the poorest

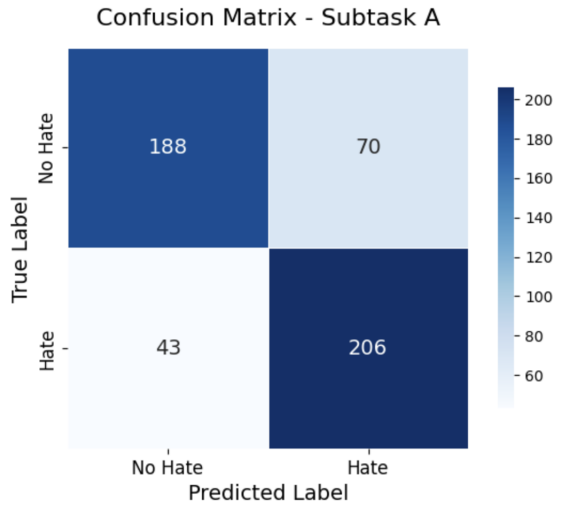

Figure 4: Confusion Matrix for Hate Speech Detection

performance (10/25, 40% accuracy), frequently confused with "Community" (10 misclassifications) and "Undirected" (3 misclassifications). This suggests the model may have difficulty in distinguishing between personal attacks and broader community-targeted content. The "Community" class achieves the best performance (82/117, 70.1% accuracy) but shows confusion with "Undirected" (21 misclassifications), indicating challenges in determining whether hate targets specific communities.

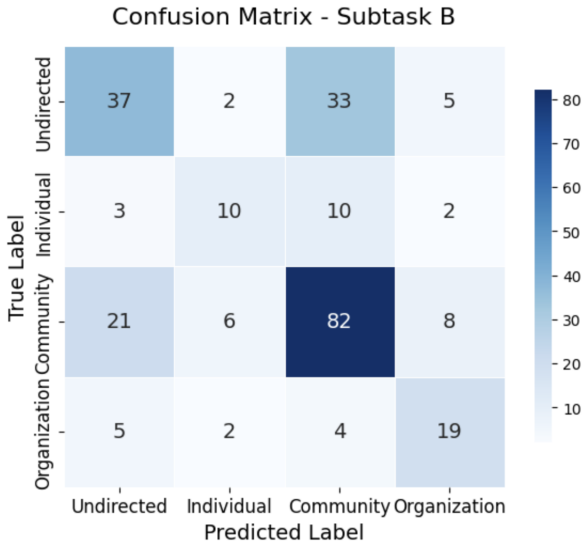

Figure 5: Confusion Matrix for Target Identification

Subtask C (Stance Classification) ensemble achieves good performance on the "Neutral" class (101/146, 69.2% accuracy) and "Oppose" class (118/169, 69.8% accuracy), but struggles with the "Support" class (98/191, 51.3% accuracy). There are 60 instances where supportive content was incorrectly classified as neutral. The model has difficulty in distinguishing between implicit support and neutral stance. Support is the most challenging class, with nearly half of supportive instances (93/191, 48.7%) being misclassified.

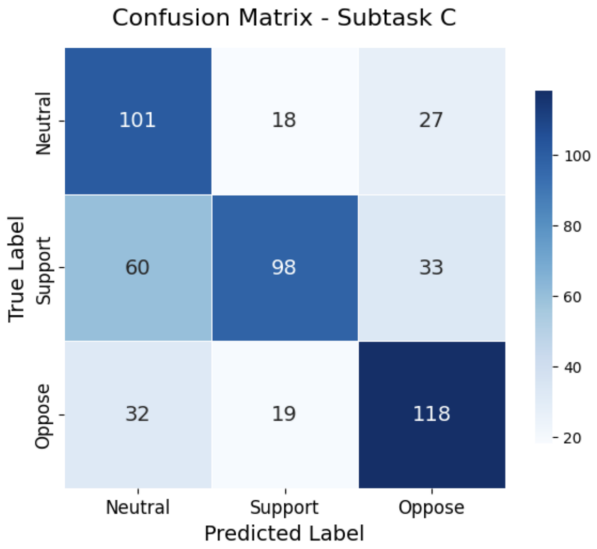

Figure 6: Confusion Matrix for Stance Classification

Subtask D (Humor Detection) shows a clear class separation. The model correctly identifies 305/342 (89.2%) humorous content and 67/165 (40.6%) non-humorous content. The error pattern reveals 98 false positives (non-humor classified as humor), suggesting the model may detect humorous elements in content intended to be serious. For example, "LGBTQ inclusive education, what conservatives think it is: here are 50 pronouns to memorize" has been incorrectly classified as Humor.

5.4 Ablation Study

To evaluate the contribution of test-time augmentation (TTA), we compared model performance with and without TTA across the subtasks. Hate speech detection showed the largest gain from F1=0.591 without TTA to F1=0.778 with TTA, while target identification improved from F1=0.510 to F1=0.553, and stance classification increased from F1=0.581 to F1=0.608. These results indicate that TTA provides significant performance benefits, with the largest improvements observed in binary classification tasks, while the more modest improvements in multi-class tasks reflects the complexity of distinguishing between fine-grained categories.

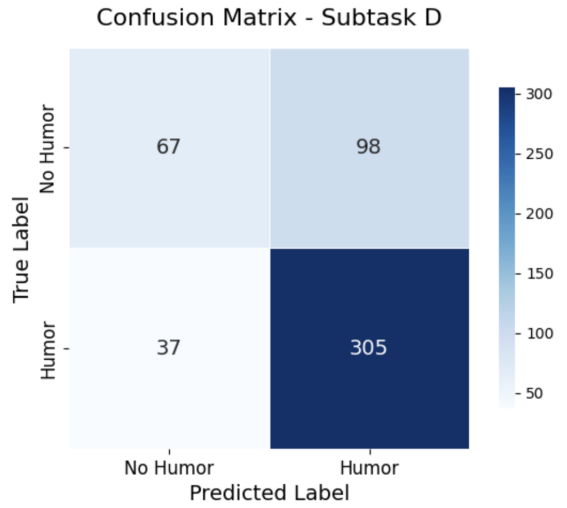

Figure 7: Confusion Matrix for Humor Recognition

6 Conclusion

In this work, we introduced a multimodal framework for Shared Task on Multimodal Hate, Humor, and Stance Detection in Marginalized Movement. We achieved F1 scores of 0.778 (hate speech detection), 0.553 (target identification), 0.608 (stance classification), and 0.658 (humor detection), which reflects the classification challenge in each of the subtasks. For hate speech detection and target identification, our RoBERTa-ResNet50 architecture with cross-modal attention performed better. While stance classification with ensemble strategies and conservative regularization, to prevent overfitting, gave us better results. Humor recognition required more advanced cross-modal attention and gating mechanisms with DialoGPT for conversational language understanding. The application of focal loss for class imbalance, test-time augmentation for robustness contributed to reliable performance across all tasks. Future work can explore ablation studies to evaluate the impact of different attention mechanisms and loss functions. Further research will focus on exploring vision-language transformers (e.g., CLIP), hierarchical attention mechanisms, and semi-supervised learning on unlabeled multimodal data.

Limitations

Some limitations emerged from our analysis that may affect the generalizability and performance of our system. First, the dataset ranges from 1,985–4,050 samples per task which can increase the risk of overfitting, particularly for deeper architectures like ResNet50 or complex attention mechanisms. This constraint may limit the model's ability to capture diverse visual and textual patterns. Techniques like semi-supervised learning could help with data scarcity. Second, annotation of humor and stance is subjective, making performance evaluation challenging for borderline cases. Additionally, the computational cost of ensemble models and cross-modal attention mechanisms restricts real-time deployment. Finally, despite using focal loss and weighted sampling, our models are sensitive to class imbalances.

References

Issa Annamoradnejad and Gohar Zoghi. 2020. Colbert: Using bert sentence embedding in parallel neural networks for computational humor. *arXiv preprint arXiv:2004.12765*.

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.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision (ECCV)*, pages 104–120. Springer.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International Conference on Web and Social Media (ICWSM)*, pages 512–515. AAAI.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

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

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Brian Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11. Association for Computational Linguistics.

Liunian Harold Li, Mark Yatskar, Da Yin, Chieh Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2920–2931. Association for Computational Linguistics.

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

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41. Association for Computational Linguistics.

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.

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. 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)*.

Yeshan Wang and Ilia Markov. 2024. CLTL@multimodal hate speech event detection 2024: The winning approach to detecting multimodal hate speech and its targets. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 73–78, St. Julians, Malta. Association for Computational Linguistics.

Yisen Wang, Zhou Xu, Chenglong Xu, and Dacheng Tao. 2019. Implicit semantic data augmentation for deep networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93. Association for Computational Linguistics.

Orion Weller and Kevin Seppi. 2020. Humor detection: A transformer gets the last laugh. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3621–3625. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 270–278.