MLInitiative at CASE 2025: Multimodal Detection of Hate Speech, Humor, and Stance using Transformers

Ashish Acharya[1*], Ankit B.K.[2*], Bikram KC[3],
Sandesh Shrestha [3], Tina Lama[3], Surabhi Adhikari[4] , Rabin Thapa[3]
[1]Kathmandu University, Dhulikhel, Nepal
[2]Thapathali Campus, Tribhuvan University, Kathmandu, Nepal
[3]IIMS College, Kathmandu, Nepal
[4]Columbia University, USA
{ankitbk75, ashishacharya048, kcvikram44, sandeshrestha115, tinal030110}@gmail.com
rabin@iimscollege.edu.np
*These authors contributed equally to this work.

Abstract

In recent years, memes have developed as popular forms of online satire and critique, artfully merging entertainment, social critique, and political discourse. On the other side, memes have also become a medium for the spread of hate speech, misinformation, and bigotry, especially towards marginalized communities, including the LGBTQ+ population. Solving this problem calls for the development of advanced multimodal systems that analyze the complex interplay between text and visuals in memes. This paper describes our work in the CASE@RANLP 2025 shared task. As a part of that task, we developed systems for hate speech detection, target identification, stance classification, and humor recognition within the text of memes. We investigate two multimodal transformer-based systems, ResNet-18 with BERT and SigLIP2, for these sub-tasks. Our results show that SigLIP-2 consistently outperforms the baseline, achieving an F1 score of 79.27 in hate speech detection, 72.88 in humor classification, and competitive performance in stance 60.59 and target detection 54.86. Through this study, we aim to contribute to the development of ethically grounded, inclusive NLP systems capable of interpreting complex sociolinguistic narratives in multi-modal content.

1 Introduction

Memes on social media are popular for their entertaining, critical, and political uses. Although memes are widely enjoyed for their entertainment value, they are increasingly exploited to propagate hate speech, circulate misinformation, and deepen prejudice, particularly against the LGBTQ+ community. For this reason, systems intended to identify and limit online toxic content must understand the diverse and many-faceted nature of memes.

New approaches in multimodal NLP now allow models to analyze both images and text within memes, which has greatly increased the accuracy of systems for detecting hate speech (Radford et al., 2021; Li et al., 2019; Velioglu and Rose, 2020). Particularly, the use of models like CLIP (Radford et al., 2021) and VisualBERT (Li et al., 2019) has opened new approaches to fuse images and texts, enhancing our understanding of their inter-modal relationships. Yet, most current methods do not address the sensitive aspects of LGBTQ+ conversations that often include personal preferences, hidden meanings, and layers of humor (Shah et al., 2024a; Thapa et al.).

Despite advancements in NLP, detecting hate speech in multimodal contexts remains challenging, primarily because images and video significantly influence message interpretation (Kiela et al., 2021). Increasingly, researchers are stressing that it's important to tell apart the targets of hate, whether they are individuals, communities, or organizations (Lee et al., 2021). These differences matter in LGBTQ+ discussions, given that the line between who individuals are and how all lesbian and gay people are seen is frequently unclear (Hardalov et al., 2022; Thapa et al., 2024).

Alongside identifying hate speech and toxicity (Rauniyar et al., 2023; Jafri et al., 2024; Naseem et al., 2025; Jafri et al., 2023), classifying people's stances is now seen as a main task for interpreting user points of view. Prior research has analyzed and detected stances in written and mixed information for uses such as political discussions and detecting falsehoods (Hasan et al., 2019; Thapa et al., 2024). At present, stance detection methods often struggle to pick up the hidden expressions of support or opposition in LGBTQ+-related memes that might involve heavy use of humor or sarcasm (Hardalov et al., 2022).

Spotting humor in memes adds more to the challenge. While hate speech makes its meaning clear, humor can hide hate underneath its jokes or sarcas-

Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025, pages 91–97, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-011

tic humor, allowing it to exist with both sides of the discussion (Shah et al., 2024a). Trying to read someone's intent can be difficult when their words are funny or sarcastic, which is why it's necessary to model humor as well as other tasks (Swamy et al., 2020).

As a consequence of the complexity of this problem, the CASE@RANLP 2025 shared task provides a complete benchmark for hate speech detection, target identification, stance assessment, and humor detection in text-embedded memes related to LGBTQ+ issues. This project aims to initiate the creation of systems that can understand sociolinguistic narratives expressed through the visual and textual interplay. focusing on LGBTQ+ issues allows the task to work toward the more general goal of increasing inclusivity and equity in the field of Natural Language Processing. This work aims to advance the field of computational social science and encourage more ethical AI within social issues and sensitive discourse using domain-specific datasets, advanced transformer models, and after applying NLP techniques to sensitive and often divisive AI. The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 outlines the dataset and task setup, Section 4 details our methodology, Section 5 presents results and discussion, and Section 6 concludes the paper.

2 Related works

Detecting hate speech in text form has gained considerable attention and has been on par with the level of human detection (Thapa et al., 2023), whereas multimodal forms like text and images are an evolving and intricate challenge due to their dynamic nature. (Kiela et al., 2020) made the Hateful Memes dataset widely available and pointed out that using information from one source only is not enough, where, through ViLBERT, they achieved only modest results. CLIP-based model has shown superior performance with HateCLIPper and MemeCLIP both achieving a macro F1 score near 0.75 on PrideMM (Shah et al., 2024a). A new study reveals that with LLM-based prompting, GPT-4 and similar LLMs come close to detecting hate speech in multimodal meme content (Zhuang et al., 2025).

Hate may be directed at different targets, so recognizing the target is also very important. Datasets such as HarMeme and PrideMM use the categories undirected, individual, community, and organization for their analysis of targets (Pramanick et al., 2021). Combining CLIP and attention brings better results than unimodals, though the overall performance is not high (F1 0.57–0.59) because the task is quite complex.

Over the years, stance detection has progressed significantly, especially with the rise of the Large Language Model. Encoder-based models like BERT & RoBRETa perform decently across data, where decoder-based models like GPT-4 showed stronger results on CLimateMist dataset (Pangtey et al., 2025). However, stance detection in memes remains unexplored. By labeling memes related to LGBTQ+ topics, PrideMM fills the gap and MemeCLIP exceeds plain text models, proving how significant this method is (Shah et al., 2024a). The same results are noticed in studies on tweets and captions: including visual and textual input together improves stance detection (Liang et al., 2024).

Recently, humor detection systems have also been improved in NLP, in addition to the simple NLP task (Shiwakoti et al., 2024). In detecting humor in text-based datasets, BERT, a transformer-based model, has shown exceptional performance, achieving an accuracy of 74% on joke assessment tasks, which was initially trained on Reddit ratings. However, it reached 98.6% accuracy on the short jokes dataset, and on the Pun of the Day dataset, it achieved the accuracy of 93.0%, outperforming another previous CNN-based model (Pangtey et al., 2025). Also, Models like MISA combined with DialogueRNN could achieve an F-score of 71.67% in multi-modal humor recognition on Hindi conversations using text, acoustic, and visual inputs, outperforming unimodal and bimodal setups, which highlights the importance of contextual and multimodal fusion in humor detection (Zhuang et al., 2025).

In addition, the development of deep learning frameworks that incorporate features like incongruity and subjectivity, along with LSTM or CNN-based textual representations, has led to the advancement in domain-specific humor detection, such as product question answering. An accuracy of 90.76% on biased datasets and 84.41% on unbiased ones has been achieved using these hybrid-type models (Zhuang et al., 2025). Recently developed architectures have been outperformed in such a type of task compared to earlier approaches and models like statistical models, N-gram analysis, and CNNs without context-specific features.

3 Dataset and Task Description

The shared task consists of four different subtasks: Subtask A aims to detect hate speech in images, Subtask B is related to classifying the targets of Hate Speech, Subtask C concentrates on categorizing images based on their stance towards the marginalized movement, and finally, Subtask D intends to detect humor in images. For all of the mentioned subtasks, datasets were provided by the organizers, which were initially created and curated by different papers (Thapa et al., 2025; Hürriyetoğlu et al., 2025; Shah et al., 2024b; Bhandari et al., 2023)

3.1 Sub-Task A

It involves binary classification to distinguish between images classified as HATE (labeled as 1) and NO-HATE (labeled as 0). For the training purpose of this sub-task, a total of 4,050 datasets were provided, out of which 1,985 were labeled as Hate and 2,065 were labeled as No Hate. The other 506 data sets were for evaluation, and 507 for testing purposes.

Figure 1: Sub-Task A Training Data Example (Left: Hate, Right: No Hate)

3.2 Sub-Task B

This sub-task is on multi-class classification for identifying the targets of hate speech. The classes contain Undirected (labeled as 0), Individual (labeled as 1), Community (labeled as 2), and Organization (labeled as 3). The associated dataset consists of 1985 training, 248 evaluation, and 249 testing datasets.

3.3 Sub-Task C

Sub-task C also involves multi-class classification for categorizing images based on their stance toward the marginalized movement. Datasets comprise 3 labels, i.e., Neutral (labeled as 0), Support (labeled as 1), and Oppose (labeled as 2), and contain 4,050 training, 506, and 507 evaluation and testing datasets, respectively.

Figure 2: Sub-Task B Training Data Example (Top-Left: Community, Top-Right: Organization, Bottom-Left: Individual, Bottom-Right: Undirected)

Figure 3: Sub-Task C Training Data Example (Left: Oppose, Middle: Neutral, Right: Support)

3.4 Sub-Task D

This last task aims to identify images showcasing humor, sarcasm, or satire related to the marginalized movement and has a binary label (i.e., no humor (labeled as 0) and humor (labeled as 1))dataset. As a dataset, 4,050 training, 506 evaluation, and 507 testing datasets were provided.

Since we don't have access to an OCR extraction tool, such as the Google Vision API, Tesseract, EasyOCR, etc., we had to use the pre-extracted text put there by the dataset organizers. In the official benchmarking paper (Shah et al., 2024a) connected with this dataset, the authors executed their OCR using the Google Vision API and trained their models in that way.

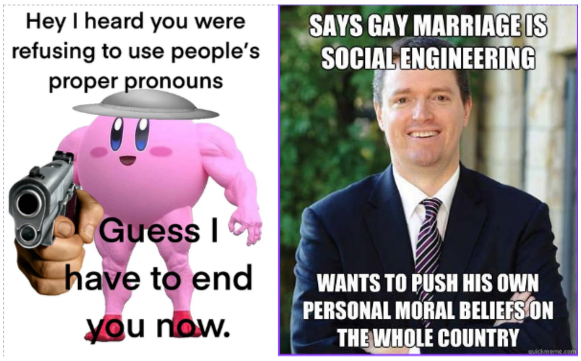

Figure 4: Sub-Task D Training Data Example (Left: No Humor, Right: Humor)

Subtask	Class	Train	Eval	Test
	Hate	1,985	248	249
A	No-Hate	2,065	258	258
	Community	931	116	177
B	Individual	199	25	25
	Organization	238	30	30
	Undirected	617	77	77
	Neutral	1166	146	146
C	Oppose	1357	169	191
	Support	1527	191	146
	Humor	2737	342	342
D	No Humor	1313	164	165

Table 1: Distribution of instances across four sub-tasks

4 Methodology

Our approach comprises two different multimodal architectures for the different sub-tasks. The first one includes ResNet-18 for the extraction of visual features and BERT-base for textual encoding, followed by a concatenation of the two feature vectors and two classification layers. The second approach uses the SigLIP2-large-patch16-256 transformer that processes image-text pairs in a joint embedding space with a sigmoid-based contrastive loss, thereby learning in a more efficient and scalable manner than CLIP.

4.1 Image Pre-processing

For the ResNet+BERT pipeline, a custom PyTorch Dataset class handled text tokenization using BERT and image preprocessing through standard ResNet transforms. For SigLIP2, the HuggingFace processor was employed to tokenize text and to transform images into model-ready tensors to maintain compatibility with the SigLIP architecture. Both loaders supported labeled and unlabeled data and returned batched inputs from a PyTorch DataLoader.

4.2 Models Architecture

4.2.1 SigLIP2

SigLIP2 is a lightweight and efficient multi-model that uses frozen image and text encoders, which means during training, no further fine-tuning can be done (Pangtey et al., 2025). The embeddings are concatenated, mapped to a lower dimension, passed through a ReLU activation function, regularized by dropout, and then classified. Due to a fixed set of feature representations, the fixed encoders allowed training to be relatively simple.

4.2.2 ResNet-18 + BERT

This multimodal architecture utilizes ResNet-18 for extracting image features, while BERT works on text embeddings(Pangtey et al., 2025). In these models, the features of visual and text are separately extracted and concatenated before moving to the next fully connected layers. Since the dataset labels were imbalanced, label smoothing was implemented in this architecture.

ResNet-18

It is a convolutional neural network with 18 layers(He et al., 2016). In this architecture, residual connections help to optimize deep networks and avoid vanishing gradient obstacles. Since it is lightweight and efficient, it is easy to classify the images using this model. It has great power to generalize over a visual task since it is pretrained on ImageNet.

BERT

Bidirectional Encoder Representations from Transformers(Devlin et al., 2019) is a language model based on a transformer-based architecture. Because it is trained on larger corpora, it is good at capturing context from words by using attention in a bidirectional manner.

4.3 Hyperparameter

Different hyperparameters were used and adjusted accordingly in two distinct model architectures for all the sub-tasks, as detailed in Table 2 below.

Parameter	Search Space	Distribution
Batch size	[16,32]	Discrete
Learning Rate	[1e-6, 5e-5]	Log-uniform
Weight Decay	[1e-6, 1e-3]	Log-uniform
Epochs	10	Discrete
Optimizer	AdamW	Categorical

Table 2: Search space for Transformer models

5 Results and Dissucssion

Table 3 compares the accuracy, recall, and F1-score of each model in all sub-tasks. Our pre-trained model, SIGLIP-2, with a fine-tuned configuration, outperforms the multimodal architecture with ResNet and BERT in each task.

Comparing F1-score of each model in the different sub-tasks. SigLIP-2 has a better score than the ResNet-18+BERT, securing 79.27%, 54.86%, 60.59% , and 72.88% , while on the other hand, ResNet-18+BERT has achieved 67.12%, 45.30%, 55.80% , and 66.45% , respectively, in Sub-Task A, Sub-Task B, Sub-Task C, and Sub-Task D.

Beyond F1-score performance, SigLIP-2 has the highest recall and accuracy of 79. 29% and 79.27% in the Hate Speech Detection task out of all tasks. However, it only achieves a recall of 58.28% while identifying targets of Hate Speech. Like SigLIP-2, ResNet-18+BERT had the highest recall of 72.86% in Sub-task 1. But, it has the lowest recall of 41.18%. Both models perform relatively poorly on Sub-task B as a result of the imbalanced dataset associated with it. In Sub-Task D, SigLIP-2 recall was able to identify humor instances, even if its overall accuracy is modest. While ResNet combined with BERT was able to accurately identify more humor and non-humor instances.

6 Limitation

Despite strong results, our models reduced performance in Sub-task B and Sub-task C, which contained imbalanced class distributions. In such situations, the model is inclined to favor the majority classes present during the training phase, showcasing its limited capability to effectively handle imbalanced instances within the datasets.

7 Conclusion

In this research, we have used two different models, one a combination of ResNet-18 + BERT, and the other is SigLIP2, for all four subtasks of the shared task. Among the used models, our model SigLIP2 performs well in all tasks. On evaluation metrics of F1-Score, this model achieves a score of 79.27% , 54.86 %, 60.59% , and 72.88 % on Sub-Task A, Sub-Task B, Sub-Task C, and Sub-Task D, respectively, which placed us 9th, 8th, 7th, and 7th, respectively, on the leaderboard for Sub-Task A, Sub-Task B, Sub-Task C, and Sub-Task D.

References

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. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A survey on stance detection for mis- and disinformation identification.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. Ur-funny: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

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

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

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

Model	Subtask	Accuracy	Recall	F1 Score
SigLIP-2	A	**0.7927**	**0.7929**	**0.7927**
	B	**0.5249**	**0.5823**	**0.5486**
	C	**0.6059**	**0.6114**	**0.6059**
	D	**0.7172**	**0.7771**	**0.7288**
ResNet-18 + BERT	A	0.7484	0.7286	0.6712
	B	0.4118	0.5040	0.4530
	C	0.5723	0.5456	0.5580
	D	0.7105	0.6311	0.6645

Table 3: Performance of deep learning models across four different tasks

Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624. Curran Associates, Inc.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2021. The hateful memes challenge: Detecting hate speech in multimodal memes.

Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. 2021. Disentangling hate in online memes. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 5138–5147. ACM.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language.

Bin Liang, Ang Li, Jingqian Zhao, Lin Gui, Min Yang, Yue Yu, Kam-Fai Wong, and Ruifeng Xu. 2024. Multi-modal stance detection: New datasets and model. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12373–12387, Bangkok, Thailand. Association for Computational Linguistics.

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.

Lata Pangtey, Anukriti Bhatnagar, Shubhi Bansal, Shahid Shafi Dar, and Nagendra Kumar. 2025. Large language models meet stance detection: A survey of tasks, methods, applications, challenges and future directions.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. 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.

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

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.

Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. 2024a. 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.

Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. 2024b. 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).

Steve Durairaj Swamy, Shubham Laddha, Basil Abdussalam, Debayan Datta, and Anupam Jamatia. 2020. NIT-agartala-NLP-team at SemEval-2020 task 8: Building multimodal classifiers to tackle Internet humor. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1179–1189, Barcelona (online). International Committee for Computational Linguistics.

Surendrabikram Thapa, Farhan Ahmad Jafri, Kritesh Rauniyar, Mehwish Nasim, and Usman Naseem. 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. 2024. 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, Shuvam Shiwakoti, Sweta Poudel, Usman Naseem, and Mehwish Nasim. 2023. 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 Sociopolitical Events from Text (CASE 2025)*.

Riza Velioglu and Jewgeni Rose. 2020. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge.

Yong Zhuang, Keyan Guo, Juan Wang, Yiheng Jing, Xiaoyang Xu, Wenzhe Yi, Mengda Yang, Bo Zhao, and Hongxin Hu. 2025. I know what you meme! understanding and detecting harmful memes with multimodal large language models.