Silver@CASE 2025: Detection of Hate Speech, Targets, Humor, and Stance in Marginalized Movement

Rohan Mainali[1*], Neha Aryal[1*], Sweta Poudel[2], Anupraj Acharya[3], Rabin Thapa[1]
[1]IIMS College, Kathmandu, Nepal
[2]Kathmandu Engineering College, Tribhuvan University, Kathmandu, Nepal
[3]Pulchowk Campus, Tribhuvan University, Kathmandu, Nepal
{rohanmainali, aryal.neha33, sweta.poudel26, anupacharya1457}@gmail.com
rabin@iimscollege.edu.np
*These authors contributed equally to this work

Abstract

Memes, a multimodal form of communication, have emerged as a popular mode of expression in online discourse, particularly among marginalized groups. With multiple meanings, memes often combine satire, irony, and nuanced language, presenting particular challenges to machines in detecting hate speech, humor, stance, and the target of hostility. This paper presents a comparison of unimodal and multimodal solutions to address all four subtasks of the CASE 2025 Shared Task on Multimodal Hate, Humor, and Stance Detection. We compare transformer-based text models (BERT, RoBERTa) with CNN-based vision models (DenseNet, EfficientNet), and multimodal fusion methods, such as CLIP. We find that multimodal systems consistently outperform the unimodal baseline, with CLIP performing the best on all subtasks with a macro F1 score of 78% in sub-task A, 56% in sub-task B, 59% in sub-task C, and 72% in sub-task D.

1 Introduction

Social networks have emerged as a platform that promotes unity by amplifying the spread of ideas in creatively diverse forms (Parihar et al., 2021). However, the proliferation of various modalities in online content has resulted in a rapid increase in hate speech, toxicity, offensive nuances, and propaganda (Rauniyar et al., 2023; Thapa et al., 2023; Jafri et al., 2024; Naseem et al., 2025; Jafri et al., 2023). A popular multimodal form of such content is memes, a combination of image or video and text that expresses ideas of a certain group or culture (Suryawanshi et al., 2020). Usually used as a powerful medium for satire, critique, and nuanced messages, memes blur the line between humor and hate, making them extremely cumbersome for machines to identify and tackle (Pramanick et al., 2021). This complication is particularly pronounced in marginalized spaces, especially the LGBTQ+ movement, where memes serve as both a means of solidarity and a force of resistance, making the content simultaneously supportive and hostile (Bikram Shah et al., 2024; Khatoon et al.).

With substantial interest from scholars and researchers, recent advances have demonstrated a significant improvement in understanding content that integrates both text and visual elements. Transformer-based models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have demonstrated strong performance, particularly when dealing with nuanced or context-dependent textual language. On the visual side, Convolutional Neural Network (CNN) architectures such as EfficientNet (Tan and Le, 2019) are widely used to extract semantic representations from images. Furthermore, CLIP (Radford et al., 2021), a vision language model trained on image-text pairs, has emerged as a powerful tool for aligning textual and visual semantics in a joint embedding space.

In this paper, we discuss a unified approach that leverages multiple deep learning techniques, including BERT, RoBERTa, DenseNet, EfficientNet, and CLIP, to detect hate speech, humor, stance, and identify target memes. These models were evaluated as part of the Shared Task on Multimodal Hate, Humor, and Stance Detection in the context of marginalized movement at CASE 2025 (Thapa et al., 2025a; Hürriyetoğlu et al., 2025). The shared task consists of four subtasks: detecting hate speech (subtask A), identifying hate targets (subtask B), classifying the topical position (subtask C), and detecting humor (subtask D) using the PrimeMM dataset associated with the LGBTQ + Pride movement (Bikram Shah et al., 2024). Our approach incorporates both unimodal and multimodal pipelines, with comparative evaluations to assess the performance and limitations of each model.

Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025, pages 83–90, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-010

2 Related Works

The increasing prevalence of using multimodal content to disseminate hate has evidently gathered the interest of researchers in developing an efficient system to detect and decrease the spread of online negativity (Gandhi et al., 2024). Research in the detection of harmful and offensive content has been conducted in both unimodal and multimodal forms, with multimodality gaining exponential popularity in recent years. Several studies on understanding the multimodality in content have demonstrated promise in addressing the challenges of harmful content on social platforms using different techniques and frameworks (Thapa et al., 2024a). More extensive research on multimodal hate detection began with a challenge organized by Facebook AI, namely the Hateful Meme Challenge, with respect to which various papers and systems have emerged to tackle this issue (Kiela et al., 2020).

2.1 Unimodal Hate Meme Detection

Traditionally, detection models were mainly based on textual content, which was later expanded to images as well. Textual models have shown a strong base with state-of-the-art performance even in noisy and nuanced language. Text-based models have been particularly dominant, employing traditional machine learning techniques such as SVMs and logistic regression with handcrafted features (Schmidt and Wiegand, 2017). They have later progressed to deep learning models, including LSTMs and transformer-based architectures like BERT (Devlin et al., 2019). These models demonstrate improved performance in identifying explicit hate speech, but struggle to capture implicit or sarcastic expressions, especially when critical context is embedded visually. Parallelly, computer vision has also advanced to provide strong performance in hate detection in images as well. Image-only models, often based on CNNs or architectures like DenseNet (Huang et al., 2017) and ResNet (He et al., 2016)— focus on visual symbolism or hateful graphics but lack the linguistic information necessary to interpret captions or textual overlays. While unimodal approaches offer simplicity and lower computational cost, several studies have shown that they are insufficient for decoding the complex interplay between text and image that characterizes modern hate memes (Kiela et al., 2020).

2.2 Multimodal Hate Meme Detection

Throughout the years, multiple efforts have been made to create a multimodal dataset of harmful and offensive memes. Most datasets focus on a specific domain or target group of hate. PrideMM dataset (Bikram Shah et al., 2024) is an annotated multimodal dataset that focuses primarily on the LGBTQ+ movement. Suryawanshi et al. (2020) suggested the MultiOFF dataset, which is related to offensive content from the 2016 US presidential election, and implemented an early fusion model to classify memes. Pramanick et al. (2021) proposed the extension of their HarMeme dataset by including US political memes as Harm-P and COVID-19 memes as Harm-C to cover larger yet specific contexts of harmful meme analysis and further annotated types of targets. A more general and nuanced dataset introduced to capture the vague sense of memes is Multi3Hate, the first multimodal and multilingual dataset with 1,500 memes, including memes in five different languages (Bui et al., 2024).

Advanced models have lately been presented that deal with the complexity of the multimodal meme. More recently, models such as CLIP (Radford et al., 2021) have bridged the gap between vision and language by demonstrating strong performance in zero-shot and few-shot classification, making it a promising model used as a base for their architecture by many researchers (Bikram Shah et al., 2024; Kapil and Ekbal, 2025). A notable research in this domain is MOMENTA, which utilized a multimodal neural network combining local and global features, and adding intramodel attentions to form the CLIP features outperforming several rivaling approaches (Pramanick et al., 2021). A unique approach was adopted in KnowMeme by leveraging a graph neural network to identify implicitly offensive content in memes with common sense (Shang et al., 2021). Recently, the use of LLMs and VLMs with zero-shot setting (Bui et al., 2024), Chain-of-Thought (Yang et al., 2023), and Chain-of-Expression (Huang et al., 2022) as well as prompting techniques (Niu et al., 2024; Sun et al., 2023), is gaining popularity in multimodal hate detection (Thapa et al., 2025b). Question-Answering has also been on the rise in the field of hate meme classification (Anaissi et al., 2025; Nandi et al., 2024).

Moreover, in the domain of humor detection, sarcasm often coexists with offensive or hateful undertones, making it a particularly challenging aspect for automated systems to detect reliably (Shiwakoti et al., 2024). Stance detection, on the other hand, has been studied in textual political and climactic discourse (Küçük and Can, 2020; Thapa et al., 2024b). However, relatively few works have tackled it in multimodal forms where visual rhetoric plays a key role. Niu et al. (2024) introduced the MmMtCSD dataset for multimodal stance detection and proposed a framework that leverages LLMs for the integration. A considerable amount of research has been conducted, particularly on the hate detection task; however, the other objectives have limited resources available in the context of multimodal content.

3 Datasets

The dataset used in this shared task is a multimodal, multi-aspect resource, PrideMM (Bikram Shah et al., 2024). The dataset comprises 5,063 text-embedded images - primarily memes - relevant to the LGBTQ + movement that are collected from Facebook, Twitter, and Reddit between 2020 and 2024. Each image in the dataset is annotated across four distinct subtasks: Hate Speech Detection, Target Classification, Topical Stance Classification, and Humor Detection. Extracted text from the text-embedded image is also provided using the OCR vision API. The dataset was segmented into train, evaluation, and test sets, with the test labels remaining undisclosed throughout the challenge. Table 1 provides the statistics of the dataset used in each of the subtasks.

Subtask	Class	Train	Eval
Subtask A	Hate	1,985	248
	No Hate	2,065	258
Subtask B	Individual	199	25
	Community	931	116
	Organization	238	30
	Undirected	617	77
Subtask C	Support	1,527	191
	Oppose	1,357	169
	Neutral	1,166	146
Subtask D	Humor	2,737	342
	No Humor	1,313	164

Table 1: Dataset Statistics of all subtasks

3.1 Subtask A: Hate Speech Detection

For sub-task A, the provided dataset contains images labeled either No Hate(0) or Hate(1), with a total of 4,050 training images and 507 images for testing. The dataset for sub-task A is quite balanced, with 1,985 instances labeled as hate, and 2,065 labeled as no hate in the provided training set. Additionally, 506 samples were also provided for evaluation, with 248 hate and 258 no-hate samples.

3.2 Subtask B: Hate Target Classification

Subtask B aligns with the classification of targets of hate speech in the text-embedded images. With a total of 1,985 memes in the training set, the targets are classified as Undirected (0) with 617 instances, Individual (1) with 199 instances, Community (2) with 931 instances, and Organization (3) with 238 instances. The evaluation set contains 248 text-embedded images, and the test set has 249 unlabeled instances.

3.3 Subtask C: Stance Classification

In Subtask C, the main objective is to determine the stance of the image, with a total of 5,063 samples annotated as Neutral (0), Support (1), or Oppose (2). The training dataset contains 1,527 samples of support, 1,357 of oppose, and 1,166 of neutral instances. Additionally, a total of 506 samples in the evaluation set contain 191 samples of support, 169 of opposition, 146 of neutral instances, and 507 images in the test set.

3.4 Subtask D: Humor Detection

Subtask D is a binary classification task focused on identifying whether the text-embedded image employs humor or not in the context of LGBTQ+ discourse. There are a total of 4,050 instances in the training set, with 2,737 labeled as humor and 1,313 labeled as no humor. The evaluation set contains 1,012 images, and the test set contains 507 images.

4 Methodology

All four subtasks have been configured with both unimodal and multimodal approaches to compare the performance of each pre-trained model for each modality. Starting with data pre-processing, model adaptation, and fusion strategies, the process and models are unified for all subtasks.

4.1 Data Processing

The multimodal nature of the dataset requires processing to be done on both the text and the image. In this section of the paper, textual and image processing, including the modeling architectures, are described. The extracted textual data obtained using the OCR technology, provided along with the dataset, was utilized for the processing of the texts. Industry standard preprocessing and normalization techniques, including lowercasing, removal of punctuations and extra whitespace, and other characters, were applied. The text was then tokenized using the HuggingFace tokenizers.

For image processing, the images were first loaded and transformed using the PIL library. Simple preprocessing steps were applied, including resizing, normalization, and data augmentation, to obtain clean and consistent data for processing. Furthermore, to ensure the alignment between the image and text for the multimodal approach, a shared index was curated with the textual data extracted from OCR.

4.2 Model Architectures

This section describes the models used in both the unimodal and multimodal settings. The unimodal approach describes both the textual and the image encoders. We utilized an extensive array of models in all subtasks to compare both unimodal and multimodal approaches. Popular transformer-based text models, BERT-base (Devlin et al., 2019) and ROBERTa-base (Liu et al., 2019), were fine-tuned to be used as the primary unimodal text models. To capture the spatial features in the image-only baselines, DenseNet-161 (Huang et al., 2017) and EfficientNet-B3 (Tan and Le, 2019) were used with the ImageNet-pretrained weights, followed by modification of the classification layer according to the number of classes in each of the subtasks. Utilizing RoBERTa-base encoder for text and EfficientNet-B3 for the images, a fusion technique was employed by concatenating the features from the two models, which achieved the best performance among multiple other combinations of text and image processing models (Habib et al., 2024). Moreover, the result was compared with the CLIP model (Radford et al., 2021) that encodes both the input modalities in a combined embedding space. The shared embeddings were trained on a custom classification head after freezing the CLIP backbone.

5 Experiments

Each model, except the frozen CLIP backbone, was fine-tuned with the AdamW optimizer with a learning rate of 1e-5 and batch size 8. All the models were trained until a maximum of 5 epochs with early stopping using the macro-averaged F1 score of the validation set. In the case of binary classification tasks, the classification threshold was also optimised based on the validation scores. All the experiments were conducted using PyTorch, text models were run on HuggingFace transformers, and images were run on timm/torchvision. Reproducibility was ensured by random seeds.

Parameter	**Value**
Learning Rate	1e-5
Batch size	8
Epochs	5
Optimizer	AdamW

Table 2: Configuration parameters

6 Result and Discussions

The performance of all models is reported using the macro F1 score, which is the official metric of the subtask. It is well-suited for this shared task due to the presence of the imbalanced classes in the subtasks. Table 3 summarizes the results of all the models implemented per subtask, reflecting the superior performance of CLIP in all subtasks. In the hate speech detection task (subtask A), multimodal models showed promising results, with the CLIP model achieving the best F1 score of 78.28%, followed by fusion of EfficientNet and RoBERTa with 76.33%. Text-based unimodal, such as RoBERTa-base, also performed quite well with an F1-score of 76.12%, presumably because the captions extracted by OCR are informal and tweet-like. Nevertheless, these models often confused sarcastic or ironic material, particularly where hate was conveyed using visual metaphors or jokes, rather than the hate being expressed through words. In contrast, image-based unimodals, EfficientNet, and DenseNet were much less effective, which validates that visual cues cannot be sufficient to effectively detect hate speech in memes.

Subtask B was particularly challenging due to the uneven distribution of the classes and the subjectivity of directed interpretation when defining the target of hateful text-embedded images. CLIP again surpassed other models with an F1 score of 56.30%, but the performance declined considerably compared to Subtask A, which suggests the complexity of the task of disambiguating the target categories. The major misclassifications were be-

Subtask	Model	Accuracy	F1 Score	Recall	Precision
Sub-Task-A	BERT-base	0.7298	0.7248	0.7276	0.7428
	RoBERTa-base	0.7613	0.7612	0.7611	0.7614
	DenseNet-161	0.6154	0.6145	0.6164	0.6179
	EfficientNet-B3	0.6291	0.6285	0.6301	0.6314
	EffNet + RoBERTa	0.7633	0.7633	0.7634	0.7633
	CLIP	**0.7830**	**0.7828**	**0.7827**	**0.7833**
Sub-Task-B	BERT-base	**0.5663**	0.5133	0.5052	0.5553
	RoBERTa-base	0.5181	0.5018	0.5422	0.5092
	DenseNet-161	0.4940	0.3859	0.3742	0.4644
	EfficientNet-B3	0.3454	0.2554	0.2745	0.2418
	EffNet + RoBERTa	**0.5663**	0.5420	0.5766	**0.5588**
	CLIP	0.5462	**0.5630**	**0.6235**	0.5421
Sub-Task-C	BERT-base	0.5680	0.5663	0.5723	0.5763
	RoBERTa-base	0.5759	0.5693	0.5709	0.5713
	DenseNet-161	0.4675	0.4570	0.4612	0.4637
	EfficientNet-B3	0.4832	0.4767	0.4777	0.4779
	EffNet + RoBERTa	0.5459	0.5393	0.5608	0.5614
	CLIP	**0.5957**	**0.5930**	**0.5947**	**0.5953**
Sub-Task-D	BERT-base	0.6923	0.6462	0.6449	0.6478
	RoBERTa-base	0.7219	0.6616	0.6543	0.6795
	DenseNet-161	0.6963	0.5275	0.6553	0.6709
	EfficientNet-B3	0.6114	0.5964	0.6195	0.6050
	EffNet + RoBERTa	0.7416	0.7053	0.7050	0.7056
	CLIP	**0.7594**	**0.7268**	**0.7275**	**0.7261**

Table 3: Performance comparison of models across subtasks A–D.

tween the groups of Community and Undirected, particularly in those memes that had broad or coded language with no explicit reference to a particular group. Also, Individual, which was the least represented category, was commonly under-predicted, even with simple upsampling used in training. This indicates a necessity for more evenly distributed training samples and possibly more detailed guidelines for annotation that would be more capable of differentiating between collective and individual targets. Both CLIP with an F1-score of 59.57% and RoBERTa at 56.93% competed well in Subtask C, which aimed to classify the stance of the meme toward marginal movements. However, when dealing with irony or tone ambiguity, even those models produced wrong classifications. The particular class of the Neutral was most likely to be miscategorized by falling into supportive or opposing messages. In addition, multimodal models, especially CLIP with an F1 score of 72.68%, performed better than the unimodal baselines in Subtask D as well, where visual cues played a major role in contextualizing comical contexts. Nevertheless, sarcasm and culturally coded jokes led to false predictions at times, especially when their models were based on images only and had no text.

7 Conclusion

In this paper, we address the multimodal and multilabel nature of the spread of online negativity using various deep learning models. We assessed the performance of each model in each sub-task and proposed a multimodal classification pipeline using CLIP to detect hate speech, classify stances, identify targets, and recognize humor in memes. By comparing transformer-based text encoders such as BERT with image encoders built on CNN architectures like EfficientNet and DenseNet, and multimodal models such as CLIP, we find that CLIP outperforms all other models. CLIP-based architecture performs particularly well in decoding context-rich content and providing better generalization across a variety of meme formats. Future work aims to account for common sense reasoning, template awareness, and temporally grounded context to make the system more consistent with human understanding. In addition, it is necessary to develop unbiased and explainable multimodal architectures that would guarantee transparency and accountability in the practical moderation of hate speech.

Limitation

Although the paper highlights recent advancements in the related objectives of hate, stance, target, and humor detection, several challenges remain unsolved. The imbalance in the dataset has limited the performance of the models as it has fewer examples to learn the features of the classes with fewer instances. The performance of different models, while demonstrating a promising result, still shows the inability to deal with ambiguous sarcasm, under-represented classes, and implicit hate speech. Dealing with these limitations is important when employing the evaluated models to accurately moderate existing hate speech in online platforms.

References

Ali Anaissi, Junaid Akram, Kunal Chaturvedi, and Ali Braytee. 2025. Detecting and understanding hateful contents in memes through captioning and visual question-answering. *ArXiv*, abs/2504.16723.

Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. 2024. Memeclip: Leveraging clip representations for multimodal meme classification. *arXiv e-prints*, pages arXiv–2409.

Minh Duc Bui, Katharina von der Wense, and Anne Lauscher. 2024. Multi3hate: Multimodal, multilingual, and multicultural hate speech detection with vision-language models. *arXiv preprint arXiv:2411.03888*.

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.

Ankita Gandhi, Param Ahir, Kinjal Adhvaryu, Pooja Shah, Ritika Lohiya, Erik Cambria, Soujanya Poria, and Amir Hussain. 2024. Hate speech detection: A comprehensive review of recent works. *Expert Systems*, 41(8):e13562.

Muhaimin Bin Habib, Md Ferdous Bin Hafiz, Niaz Ashraf Khan, and Sohrab Hossain. 2024. Multimodal sentiment analysis using deep learning fusion techniques and transformers. *International Journal of Advanced Computer Science & Applications*, 15(6).

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.

Fan Huang, Haewoon Kwak, and Jisun An. 2022. Chain of explanation: New prompting method to generate quality natural language explanation for implicit hate speech. *Companion Proceedings of the ACM Web Conference 2023*.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

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)*.

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*.

Prashant Kapil and Asif Ekbal. 2025. A transformer based multi task learning approach to multimodal hate speech detection. *Natural Language Processing Journal*, 11:100133.

Kashifa Khatoon, Sania Yaseen, and Zafar Iqbal. Humor in hostility: A critical multimodal analysis of memes circulating on social media after the pehalgam attack.

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.

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*.

Palash Nandi, Shivam Sharma, and Tanmoy Chakraborty. 2024. Safe-meme: Structured reasoning framework for robust hate speech detection in memes. *ArXiv*, abs/2412.20541.

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.

Fuqiang Niu, Zebang Cheng, Xianghua Fu, Xiaojiang Peng, Genan Dai, Yin Chen, Hu Huang, and Bowen Zhang. 2024. Multimodal multi-turn conversation stance detection: A challenge dataset and effective model. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 3867–3876.

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.

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. *arXiv preprint arXiv:2109.05184*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

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.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

Lanyu Shang, Christina Youn, Yuheng Zha, Yang Zhang, and Dong Wang. 2021. Knowmeme: A knowledge-enriched graph neural network solution to offensive meme detection. In *2021 IEEE 17th International Conference on eScience (eScience)*, pages 186–195. IEEE.

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).

Xiangguo Sun, Jiawen Zhang, Xixi Wu, Hong Cheng, Yun Xiong, and Jia Li. 2023. Graph prompt learning: A comprehensive survey and beyond. *arXiv preprint arXiv:2311.16534*.

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.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

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, 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. 2025a. Multimodal hate, humor, and stance event detection in marginalized sociopolitical movements. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Sociopolitical Events from Text (CASE 2025)*.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. 2025b. Large language models (llm) in computational social science: prospects, current state, and challenges. *Social Network Analysis and Mining*, 15(1):1–30.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se young Yun. 2023. Hare: Explainable hate speech detection with step-by-step reasoning. *ArXiv*, abs/2311.00321.