Multimodal Kathmandu@CASE 2025: Task-Specific Adaptation of Multimodal Transformers for Hate, Stance, and Humor Detection

Sujal Maharjan[1*], Astha Shrestha[1*], Shuvam Thakur[2], Rabin Thapa[1]
[1] IIMS College, Kathmandu, Nepal
[2] Delhi Technological University, New Delhi, India
{sujalmaharjan007, aasthashrestha688}@gmail.com
shuvamthakur@outlook.com
rabin@iimscollege.edu.np
*These authors contributed equally to this work

Abstract

The multimodal ambiguity of text-embedded images (memes), particularly those pertaining to marginalized communities, presents a significant challenge for natural language and vision processing. The subtle interaction between text, image, and cultural context makes it challenging to develop robust moderation tools. This paper tackles this challenge across four key tasks: (A) Hate Speech Detection, (B) Hate Target Classification, (C) Topical Stance Classification, and (D) Intended Humor Detection. We demonstrate that the nuances of these tasks demand a departure from a 'one-size-fits-all' approach. Our central contribution is a task-specific methodology, where we align model architecture with the specific challenges of each task, all built upon a common CLIP-ViT backbone. Our results illustrate the strong performance of this task-specific approach, with multiple architectures excelling at each task. For Hate Speech Detection (Task A), the Co-Attention Ensemble model achieved a top F1-score of 0.7929; for Hate Target Classification (Task B), our Hierarchical Cross-Attention Transformer achieved an F1-score of 0.5777; and for Stance (Task C) and Humor Detection (Task D), our Two-Stage Multiplicative Fusion Framework yielded leading F1-scores of 0.6070 and 0.7529, respectively. Beyond raw results, we also provide detailed error analyses, including confusion matrices, to reveal weaknesses driven by multimodal ambiguity and class imbalance. Ultimately, this work provides a blueprint for the community, establishing that optimal performance in multimodal analysis is achieved not by a single superior model, but through the customized design of specialized solutions, supported by empirical validation of key methodological choices.

1 Introduction

Social media has significantly influenced public discourse, with text-embedded images, or memes, now serving as a dominant means for debate, specifically addressing the surrounding social movements and marginalized communities (Burbi et al., 2023; Thapa et al., 2024a). These multimodal artifacts reflect a broad spectrum of messages, from solidarity and support to targeted persecution and hate (Kumar and Pranesh, 2021). This dynamic is particularly evident in content relevant to the LGBTQ+ community, where memes appear as intricate instruments of in-group humor, political commentary, and nefarious attack, often simultaneously (Arcila-Calderón et al., 2021).

The multimodal ambiguity of these artifacts is the key challenge for automated analysis. The meaning of a meme is often inferred from a subtle interaction between its visual and textual components, necessitating a thorough understanding of cultural and contextual differences to interpret accurately (Kiela et al., 2020). Consequently, the line between satire and genuine offense becomes perilously unclear, presenting a substantial barrier for content moderation systems (Chavez and Prado, 2023; Naseem et al., 2025). This ambiguity highlights the constraints of simple binary classifications (e.g., hate/no hate), which fail to capture the multifaceted traits of the expression (Carvalho et al., 2024). An extensive study is therefore paramount to evaluate the entire communicative act, including its intended humor, intended targets, and overall stance.

To address this challenge, and as part of the *Shared Task on Multimodal Hate, Humor, and Stance Detection in Marginalized Movement@CASE2025* (Thapa et al., 2025), this paper presents a fine-grained, multi-task framework for the in-depth analysis of memes from the PrideMM dataset (Shah et al., 2024) related to marginalized communities, held at the 8th Workshop on Challenges and Applications of Automated Ex-

Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025, pages 107–114, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-013

traction of Socio-political Events (CASE 2025) (Hürriyetoğlu et al., 2025). Our framework concurrently addresses the four different but interrelated sub-tasks as defined by the task organizers: (A) Hate Speech Detection, (B) Hate Target Classification, (C) Topical Stance Classification, and (D) Intended Humor Detection. By tackling these aspects simultaneously, our work transcends beyond simplistic labels to offer a more enhanced and pertinent model of online multimodal communication. This research not only advances a robust system for a critical shared task but also contributes to the overarching goal of developing more accessible and efficient AI for comprehending the intricate nature of human expression online. Our tailored approach proved highly effective, securing a top-three finish in the nuanced challenge of Intended Humor Detection (Subtask D), which required identifying not just humor but also satire and sarcasm, while achieving competitive performance across all sub-tasks.

2 Related works

The task of automatically recognizing hate speech has progressed significantly, with research shifting from purely textual analysis (Rauniyar et al., 2023; Thapa et al., 2024b, 2023b; Jafri et al., 2024, 2023) to the more complicated domain of multimodal content (Baltrušaitis et al., 2018), a field encompassing a wide range of applications and challenges (Parihar et al., 2021). The growth of memes, where meaning originates through a synthetic and often non-literal interaction of image and text, has produced many text-only models that were inadequate. Kiela et al. (2020) introduced the Hateful Meme Challenge, highlighting a significant turning point for the field. It presented a carefully assembled dataset where innocuous images or text could become hateful when paired together, showing that models must engage in true multimodal thinking to succeed. This spawned the development of higher-level architectures aimed at integrating the data across various modalities. Researchers have studied numerous fusion approaches, from basic feature concatenation to more intricate co-attention approaches and dedicated fusion models like MemeFier (Koutlis et al., 2023), which uses a dual-stage technique to align and fuse the visual and textual elements.

Concurrent with the initiatives to enhance detection accuracy, substantial research inspiration has focused on achieving a detailed understanding of harmful content. Researchers began to work on finding who is being targeted after realizing that recognizing binary hate/no-hate classification alone is not sufficient. This spurred the development of datasets with multi-aspect annotations (Thapa et al., 2024c, 2023a), which not just identify the presence of hate but also its particular target attributes, such as religion, gender, or origin (Ousidhoum et al., 2019) and even whether the hate is directed or undirected (Bhandari et al., 2023). This has been further refined by more recent benchmarks such as the THOS dataset (Almohaimeed et al., 2023) by offering hierarchical labels that differentiate between general hate concerns and specific targets. This fine-grained approach also extends to stance detection, which analyzes an author's viewpoint (e.g., support, oppose) towards a specific topic or entity. This has been successfully employed in the analysis of discourse around social movements such as Black Lives Matter (Kumar and Pranesh, 2021), providing a strong methodological foundation for our subtask of classifying stance towards the LGBTQ+ community.

Perhaps the most subtle challenge lies in interpreting humor and satire, which can be used to deliver offensive messages while maintaining plausible deniability. Humor is a multifaceted social phenomenon; it can act as a key means for in-group solidarity and resilience within marginalized communities (Baker et al., 2020; Shiwakoti et al., 2024); however, it can also be used to regularize prejudice and mock hate victims (Chavez and Prado, 2023). This underlying ambiguity makes it a tremendous problem for computational systems. In response, dedicated shared tasks and datasets like MAMI (Qu et al., 2022; Hee et al., 2023) have been developed to offer a research platform for the multimodal analysis of memes, with distinguished tracks for identifying humor, sarcasm, and offence. Our work directly complements this effort by treating Humor Detection as a distinct analytical dimension, enabling us to distinguish comedic intent from hateful expression and authorial stance. By incorporating research threads such as multimodal hate detection, fine-grained target and stance analysis, and humor detection, our project aims to create a comprehensive framework for analyzing nuanced online content relevant to the LGBTQ+ community.

3 Dataset and Task

Our experiments were conducted on the PrideMM dataset (Shah et al., 2024), which was provided by the shared task organizers for this challenge. The task includes 4 different subtasks: Sub-Task A: Detection of Hate Speech, Sub-Task B: Classifying the Target of Hate Speech, Sub-Task C: Classification of Topical Stance, and Sub-Task D: Detection of Intended Humor.

3.1 Sub-Task A

This subtask is a binary classification focused on identifying hate speech. The goal is to distinguish between the content that contains hate and the content that does not contain hate. The provided dataset consists of 4,050 training samples with 1,985 samples of 'Hate' and 2,065 samples of 'NO Hate.' The number of validation samples is 506, and the number of test samples is 507.

3.2 Sub-Task B

This sub-task B focuses on classifying the target of content among 'Community', 'Individual', 'Organisation', and 'Undirected'. The training dataset consists of 1,385 samples, with 'Community' being the most frequent category with 931 instances, while the least frequent, 'Individual‘, has 199 instances. The dataset also consists of 248 validation samples and 249 test samples.

3.3 Sub-Task C

This sub-task C involves multi-class classification to determine the stance towards the given target with three labels: 'Support', 'Oppose', and 'Neutral'. The dataset consists of 4,050 training samples, with the majority, 1,527 samples, being 'support' labels. The dataset also contains 506 validation samples and 507 test samples.

3.4 Sub-Task D

The last sub-task D is also a binary classification to identify the presence of Humor. The dataset consists of 4,050 training samples with 2,737 samples of 'Humor' and 1,313 samples of 'no Humor' labels. The validation sample and test sample is consistent with sub-task A and C, containing 506 and 507, respectively.

Subtask	Class	Train	Eval	Test
A	Hate	1,985	248	507
	No Hate	2,065	258	
B	Individual	199	25	249
	Community	931	116	
	Organization	238	30	
	Undirected	617	77	
C	Support	1,527	191	507
	Oppose	1,357	169	
	Neutral	1,166	146	
D	Humor	2,737	342	507
	No Humor	1,313	164	

Table 1: Summary of Dataset Statistics

4 Methodology

Our methodology is built on task-specific adaptation. Recognizing the subtle challenges of hate speech, target identification, stance, and humor detection are not amenable to a comprehensive technique; therefore, we developed and analyzed a suite of tailored systems. This section details the architectures, fusion mechanisms, and advanced training protocols that yielded the model that performed best for each task.

4.1 Common Setup

Our systems are built upon the Contrastive Language–Image Pre-training (CLIP) family of models (Radford et al., 2021), with openai/clip-vit-large-patch14 as our primary model. At the same time, our comparative experiments for Subtask C also included the laion/CLIP-ViT-L-14-DataComp.XL-s13B-b90K model to assess scaling effects. The dataset presents several challenges, including a moderate class imbalance, which we mitigated by employing balanced class weighting within the cross-entropy loss function. Furthermore, we used a strong data augmentation strategy including Random Resized Crop (RRC) (from TorchVision) and RandAugment (Cubuk et al., 2019) to improve model invariance and dynamically handled any corrupt image files to maintain training stability. To ensure reproducibility, all single-model experiments used a fixed random seed of 42, while our ensemble for Subtask A used five unique fixed seeds.

4.2 Task-Specific Architectures

Our central hypothesis was that each subtask demands a unique modeling of the image-text interaction. For the high-variance task of Hate Speech

(A), we reasoned that an ensemble would be most effective at reducing prediction variance. For fine-grained Target Classification (B), a hierarchical attention model was developed to learn direct links between textual tokens and visual regions. In our initial experiments for Subtask B, we tested a simpler baseline using direct feature concatenation of the image and text embeddings. This approach yielded a significantly lower F1-score (0.5506 on the validation set), confirming our hypothesis that an explicit cross-attention mechanism is essential for grounding textual targets within the visual context. For Stance and Humor (C, D), which often depend on conditional interactions, we employed a multiplicative fusion framework to explicitly model this non-linear dynamic. We detail these three core architectures below.

Ensemble of Transformer-based Fusion Models (Subtask A): This architecture operates on pre-computed 768-dimensional CLIP features. For each meme, the image and text vectors are concatenated and processed by a four-layer, eight-head Transformer encoder. The final prediction is a robust average of the softmax probabilities from an ensemble of size 5, a variance-reduction technique analyzed by Andrew and Gao (2007). We chose this approach because hate speech detection is a high-variance task where subtle cues can significantly alter the classification. Ensembling helps in stabilized predictions and decrease the risk of overfitting to erroneous correlations in the training data.

Hierarchical Cross-Attention Transformer (Subtask B): This end-to-end architecture refines 768-dimensional image and text features in parallel using separate 2-layer Transformer encoders. A cross-attention mechanism then allows the textual representation to contextually query the visual representation. This contextualized text feature is then concatenated with the original refined text feature for final classification. This architecture is particularly designed for target classification as it enables the model to ground textual targets (such as 'individual', 'community') in the visual content of meme, which is critical for accurate identification.

Two-Stage Multiplicative Fusion Framework (Subtasks C & D): Inspired by the MemeCLIP approach (Shah et al., 2024), this framework first projects the 768-dimensional CLIP features into a 1024-dimensional space. These projected features are then refined using lightweight adapter modules, and their interaction is modeled through element-wise multiplication.This approach works well for the tasks like stance and humor detection, as these tasks often rely on subjective and non-linear interactions between the text and image. These intricate relationships are better captured by multiplicative fusion than by simpler additive or concatenative techniques.

Our training protocol was defined by three core techniques, with final hyperparameters Table 2 selected from a limited random search of approximately 20 trials. The empirical impact of these techniques on the validation set is shown in Table 3.

Two-Stage Fine-Tuning: This protocol was critical for the stability of our end-to-end models. In Stage 1, we froze the CLIP backbone and trained only the task-specific modules for 5–8 epochs. In Stage 2, we performed a gentle, end-to-end fine-tuning, unfreezing the final 2 layers of the CLIP encoders for up to 20 additional epochs with early stopping. This approach yielded a +2.58 F1 point gain over a frozen-backbone baseline on Subtask C. This two-stage protocol is critical for preventing 'catastrophic forgetting', where end-to-end fine-tuning can degrade the powerful, general-purpose features of the pre-trained CLIP backbone. By first training only the task-specific modules, we anchor the model in the correct feature space before gently refining the entire architecture.

Advanced Regularization and Initialization: A cornerstone of our framework for Subtasks C and D was Semantic-Aware Initialization (SAI), a technique where a Cosine Classifier's weights are seeded using CLIP-encoded embeddings of class-descriptive prompts (e.g., "a meme expressing a 'support' stance"), which consistently provided faster, more stable convergence. We also explored Stochastic Weight Averaging (SWA) (Izmailov et al., 2019) on multiple subtasks. For Subtask C, it was integral to the training process, though the final checkpoint selected was the standard (non-averaged) model which achieved the highest validation score. We note that while SWA provided a performance lift on some tasks, our task-specific 'Hierarchical Cross-Attention Transformer' for Subtask B ultimately outperformed our SWA-enhanced baselines on the validation set, suggesting that

Subtask	System Architecture	Learning Rate (Head / Backbone)	Batch Size (Effective)	Weight Decay
A	Co-Attention Ensemble	2e-4 / —[1]	1024	0.1
B	Hierarchical Cross-Attention Transformer	2e-5 / 2e-6	64	0.1
C	Two-Stage Multiplicative Fusion	2e-5 / 1e-8	16	1e-2
D	Two-Stage Multiplicative Fusion	1e-4 / 1e-6	32	1e-2

Table 2: Key Hyperparameters for Our Best-Performing Models. [1] indicates that the model was not fine-tuned.

Technique Comparison	Subtask	ΔF1 (pts)	Purpose
Two-Stage Fine-Tuning vs. a Frozen Backbone	C (Stance)	+2.58	Improves training stability
Two-Stage Fine-Tuning with SWA vs. without SWA	C (Stance)	+0.47	Smooths the optimization landscape
Ensemble of size 5 vs. the Best Single Model	A (Hate Speech)	+0.41	Reduces prediction variance

Table 3: Empirical Validation of Key Methodological Choices (on the Official Validation Set).

for this specific task, architectural innovation was more impactful than optimization smoothing. All end-to-end models employed Automatic Mixed-Precision (AMP) via `torch.cuda.amp` to accelerate training.

4.3 Implementation

All experiments were run on Google Colaboratory with a single NVIDIA T4 GPU. Automatic Mixed Precision (AMP) via `torch.cuda.amp` was used in all training runs to reduce memory usage and speed up convergence. To ensure full reproducibility and to facilitate future research, we publicly release our implementation, including code, training scripts, and the final model weights: https://github.com/SUJAL390/CASE-2025-Multimodal-Meme-Analysis.

5 Result and Discussion

Our comprehensive analysis across all four subtasks, shown in Table 4 illustrates that achieving optimal performance is accomplished by integrating specialized approaches with the unique demands of each task rather than depending on a single, universal model.Presenting the superiority of model aggregation for robust classification, a co-attention ensemble proved to be most effective for hate speech detection (subtask A), achieving a final test F1 score of 0.7929. On the other hand, the fine-grained challenge of Target Classification (Subtask B) was best addressed by architectural innovation, with the Hierarchical Cross-Attention Transformer achieving the highest F1 score of 0.5777. For Stance and Humor Detection (Subtasks C and D), superior results were achieved via advanced optimization, with Two-Stage Fine-Tuning techniques achieving the leading F1 scores of 0.6070 and 0.7529, respectively, highlighting the importance of methodical adaptation of large pre-trained models.

6 Error Analysis

The confusion matrix, presented in Figure 1, shows both the true and predicted labels, implying that our model shows a relatively balanced performance between the classes rather than a strong bias towards one. The critical errors are the 45 instances where 'Hate' was mislabelled as 'No Hate' in sub-task A. Since our training dataset is well-balanced, this issue does not trigger from data prevalence. Instead, the errors are likely to originate from the 'multimodal ambiguity' central to our paper, where complex irony or satire masks the content's true hateful intent from the model.

In sub-task B, our model is assigned with the challenge of categorizing targets from text-embedded images into four labels: 'Individual', 'Community', 'Organization', and 'Undirected.' Analysis of the confusion matrix in Figure 1 shows that our model has difficulties in identifying 'Undirected' targets, which are commonly misclassified as 'Community' (35 instances). The observed challenges in the model's performance, especially in differentiating between these two classes, can be the cause of a significant imbalance in the training dataset, as shown in Table 1.

For this sub-task C, the model is assigned to classify the stance as 'Neutral', 'Support', or 'Oppose'.

Subtask	System Architecture	Accuracy	Precision	Recall	F1 Score	Rank
A	Ensemble of Transformer-based Fusion Models	0.7929	0.7933	0.7932	**0.7929**	7
B	Hierarchical Cross-Attention Transformer	0.5823	0.5666	0.5922	**0.5777**	5
C	Two-Stage Multiplicative Fusion	0.6114	0.6218	0.6125	**0.6070**	6
D	Two-Stage Multiplicative Fusion	0.7791	0.7491	0.7578	**0.7529**	3

Table 4: Official performance of our final systems on the blind test set. For each subtask, the rank is determined by the F1 score (bold). All scores are as reported by the task organizers.

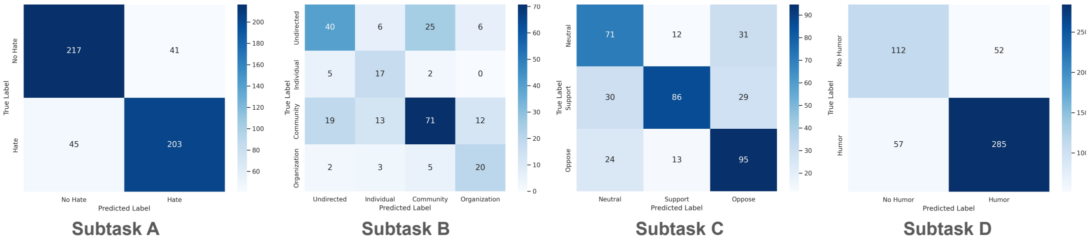

Figure 1: Confusion matrices of Subtasks A, B, C, and D on the evaluation set

The confusion matrix in Figure 1 indicates that the model most significantly struggles with the 'Neutral' class, often mislabelling it as 'Oppose' (31 instances). Moreover, a high degree of confusion exists between the 'Support' and 'Oppose' categories (29 misclassifications). This pattern highlights the challenge of assessing subjective content. The error implies that the model fails to properly comprehend sarcasm or nuanced political commentary, where the literal text and image may not align with the author's actual stance.

In the sub-task D, our aim is to perform a binary classification of 'No Humor' and 'Humor'. The confusion matrix in Figure 1 indicates that our model performs exceptionally well in recognizing 'Humor' (285 True Positives) but is significantly less accurate when dealing with 'No Humor' content (52 False Positives). The apparent bias towards recognizing humor forms in the model may arise from the substantial number of Humor-labelled texts in the training dataset, which includes more than twice as many samples as the 'No Humor' class (2,737 vs. 1,313). Since both the training and evaluation datasets are utilized to train the model, the model may develop bias, affecting its accuracy when handling the non-humorous speeches.

7 Conclusion

This research challenges the notion of a universal model for multimodal NLP. Through a rigorous, task-by-task analysis, we have demonstrated that optimal performance is not a matter of finding a single, superior architecture but of meticulously aligning specialized models with the unique demands of each task.

Our findings offer a clear blueprint for researchers: ensemble models provide the necessary stability for high-variance tasks like hate speech detection; hierarchical attention is crucial for grounding fine-grained targets; and multiplicative fusion with semantic initializations best suited for subjective interpretation tasks like stance and humor. By advocating for this paradigm shift away from a 'one-size-fits-all' approach, our work establishes that the future of high-performance, responsible NLP lies in the customized design of tailored solutions that achieve a state of task-model resonance.

8 Limitations

The underlying dataset and model design impose limitations on our shared-task submission. First, significant class imbalance and the subjectivity intrinsic in categorizing nuanced phenomena (humor, stance, hate) introduce noise that can be skewed towards dominant classes, limiting generalization to out-of-domain datasets and different cultural or linguistic contexts, as our training is based on a static snapshot of online discourse. Second, our systems may struggle with emerging meme templates and novel cultural references, or non-Western contexts which challenges the static models. Third, our top-performing ensemble architecture, while effective, is computationally expensive and difficult to interpret, limiting its deployability without further model compression or knowledge distillation. Eliminating these issues via improved sampling strategies, multilingual foundation models, and continuous learning pipelines will be critical for robust,equitable and sustainable performance.

Ethics Statement

This work follows the ACL Ethics Policy, using an anonymized dataset to develop models for detecting harmful content. While aiming to create safer online spaces, we acknowledge the potential for misuse in surveillance or censorship. To mitigate this, we have implemented fairness checks, recommend human-in-the-loop oversight for deployment, and advocate for transparent documentation and community engagement.

References

Saad Almohaimeed, Saleh Almohaimeed, Ashfaq Ali Shafin, Bogdan Carbunar, and Ladislau Bölöni. 2023. Thos: A benchmark dataset for targeted hate and offensive speech. *arXiv preprint arXiv:2311.06446*.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Carlos Arcila-Calderón, Javier J Amores, Patricia Sánchez-Holgado, and David Blanco-Herrero. 2021. Using shallow and deep learning to automatically detect hate motivated by gender and sexual orientation on twitter in spanish. *Multimodal technologies and interaction*, 5(10):63.

James E Baker, Kelly A Clancy, and Benjamin Clancy. 2020. Putin as gay icon? memes as a tactic in russian lgbt+ activism. *LGBTQ+ activism in Central and Eastern Europe: Resistance, representation and identity*, pages 209–233.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

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.

Giovanni Burbi, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. 2023. Mapping memes to words for multimodal hateful meme classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2832–2836.

Paula Carvalho, Danielle Caled, Cláudia Silva, Fernando Batista, and Ricardo Ribeiro. 2024. The expression of hate speech against afro-descendant, roma, and lgbtq+ communities in youtube comments. *Journal of Language Aggression and Conflict*, 12(2):171–206.

Jason V Chavez and RTD Prado. 2023. Discourse analysis on online gender-based humor: Markers of normalization, tolerance, and lens of inequality. In *Forum for Linguistic Studies*, volume 5, pages 55–71.

Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. 2019. Randaugment: Practical automated data augmentation with a reduced search space.

Ming Shan Hee, Wen-Haw Chong, and Roy Ka-Wei Lee. 2023. Decoding the underlying meaning of multimodal hateful memes. *arXiv preprint arXiv:2305.17678*.

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)*.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. 2019. Averaging weights leads to wider optima and better generalization.

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.

Christos Koutlis, Manos Schinas, and Symeon Papadopoulos. 2023. Memefier: Dual-stage modality fusion for image meme classification. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 586–591.

Sumit Kumar and Raj Ratn Pranesh. 2021. Tweetblm: A hate speech dataset and analysis of black lives matter-related microblogs on twitter. *arXiv preprint arXiv:2108.12521*.

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

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

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.

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

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.

Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. 2024. Memeclip: Leveraging clip representations for multimodal meme classification. *arXiv preprint arXiv:2409.14703*.

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).

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)*.