CASE 2025

Proceedings
of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts

associated with

The 15th International Conference on Recent Advances in Natural Language Processing RANLP'2025

Edited by Ali Hürriyetoğlu, Hristo Tanev, Surendrabikram Thapa, Virginia Tech and Surabhi Adhikari

13 September, 2025
Varna, Bulgaria

The 8th Workshop on Challenges and Applications
of Automated Extraction of Socio-political Events from Texts
Associated with the International Conference
Recent Advances in Natural Language Processing
RANLP'2025

PROCEEDINGS

Varna, Bulgaria
13 September 2025

Online ISBN 978-954-452-099-1

Designed by INCOMA Ltd.
Shoumen, BULGARIA

Message from the CASE 2025 Organizing Committee

Continuing its tradition, CASE brings together researchers from computational and social sciences to explore the evolving landscape of event extraction. Alongside text-based approaches, the workshop also highlights the growing interest in multimodal event extraction, addressing complex real-world scenarios across diverse modalities.

This 8th edition of the workshop underlines the increasing importance of the LLM and the deep learning architectures. With a keynote speech on LLMs and a shared task on multimodal hate, humor, and stance detection, this year workshop successfully charts the future directions in the challenging area of event detection and extraction!

Organizing Committee

Ali Hürriyetoğlu, Wageningen University & Research
Erdem Yörük, Koç University
Hristo Tanev, European Commission, Joint Research Centre
Surendrabikram Thapa, Virginia Tech
Vanni Zavarella, University of Cagliari, Italy

Volume Editors

Ali Hürriyetoğlu, Wageningen University & Research
Hristo Tanev, European Commission, Joint Research Centre
Surendrabikram Thapa, Virginia Tech
Surabhi Adhikari, Columbia University

Table of Contents

Conference Program

Findings and Insights from the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text
Ali Hurriyetoglu, Surendrabikram Thapa, Hristo Tanev and Surabhi Adhikari

Challenges and Applications of Automated Extraction of Socio-political Events at the age of Large Language Models
Surendrabikram Thapa, Surabhi Adhikari, Hristo Tanev and Ali Hurriyetoglu

Multimodal Hate, Humor, and Stance Event Detection in Marginalized Sociopolitical Movements
Surendrabikram Thapa, Siddhant Bikram Shah, Kritesh Rauniyar, Shuvam Shiwakoti, Surabhi Adhikari, Hariram Veeramani, Kristina T. Johnson, Ali Hurriyetoglu, Hristo Tanev and Usman Naseem

Natural Language Processing vs Large Language Models: this is the end of the world as we know it, and I feel fine
Bertrand De Longueville

Machine Translation in the AI Era: Comparing previous methods of machine translation with large language models
William Jock Boyd and Ruslan Mitkov

Steering Towards Fairness: Mitigating Political Stance Bias in LLMs
Afrozah Nadeem, Mark Dras and Usman Naseem

wangkongqiang@CASE 2025: Detection and Classifying Language and Targets of Hate Speech using Auxiliary Text Supervised Learning
wang kongqiang and Zhang Peng

Luminaries@CASE 2025: Multimodal Hate Speech, Target, Stance and Humor Detection using ALBERT and Classical Models
Akshay Esackimuthu

Overfitters@CASE2025: Multimodal Hate Speech Analysis Using BERT and RESNET
Bidhan Chandra Bhattarai, Dipshan Pokhrel, Ishan Maharjan and Rabin Thapa

Silver@CASE2025: Detection of Hate Speech, Targets, Humor, and Stance in Marginalized Movement
Rohan Mainali, Neha Aryal, Sweta Poudel, Anupraj Acharya and Rabin Thapa

MLInitiative at CASE 2025: Multimodal Detection of Hate Speech, Humor,and Stance using Transformers
Ashish Acharya, Ankit BK, Bikram K.C., Surabhi Adhikari, Rabin Thapa, Sandesh Shrestha and Tina Lama

Multimodal Deep Learning for Detection of Hate, Humor, and Stance in Social Discourse on Marginalized Communities
Durgesh Verma and Abhinav Kumar

Findings and Insights from the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text

Ali Hürriyetoğlu[1], Surendrabikram Thapa[2], Hristo Tanev[3], Surabhi Adhikari[4]
[1]Wageningen Food Safety Research, Netherlands, [2]Virginia Tech, USA,
[3]European Commission, Joint Research Centre, Italy, [4]Columbia University, USA
[1]ali.hurriyetoglu@wur.nl, [2]sbt@vt.edu,
[3]hristo.tanev@ec.europa.eu, [4]surabhi.adhikari@columbia.edu

Abstract

This paper presents an overview of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE), held in conjunction with RANLP 2025. The workshop featured a range of contributions, including regular research papers, system descriptions from shared task participants, and an overview paper on shared task outcomes. Continuing its tradition, CASE brings together researchers from computational and social sciences to explore the evolving landscape of event extraction. With the rapid advancement of large language models (LLMs), this year's edition placed particular emphasis on their application to socio-political event extraction. Alongside text-based approaches, the workshop also highlighted the growing interest in multimodal event extraction, addressing complex real-world scenarios across diverse modalities.

1 Introduction

In an increasingly interconnected and digitized world, the vast availability of textual and multimodal data related to socio-political, economic, environmental, and humanitarian events presents unprecedented opportunities for data-driven analysis across the social sciences and humanities (Hürriyetoğlu et al., 2024, 2021a; Chen et al., 2023). Governments, international organizations, journalists, and civil society actors increasingly rely on such data to gain timely, granular, and actionable insights into events such as protests, conflicts, public health emergencies, migration patterns, and policy shifts (Shu and Ye, 2023; Hürriyetoğlu et al., 2021c).

Recent years have seen a dramatic shift in the landscape of event extraction due to the rapid advancement of large language models (LLMs) (Thapa et al., 2025c; Hou and Huang, 2025). These models, capable of understanding, generating, and reasoning over text with minimal supervision, have opened new avenues for tackling long-standing challenges in socio-political event extraction such as low-resource languages, implicit events, cross-document reasoning, and multilingual understanding (Ziems et al., 2024; Anthis et al.; Shen et al., 2023). Techniques such as prompt-based learning, instruction tuning, and alignment have enabled more adaptive and generalizable extraction pipelines, reducing dependence on handcrafted features and rigid annotation schemas (Hou et al., 2024; Kirk et al., 2024; Khan et al., 2025).

Beyond text, the field is increasingly moving toward multimodal event extraction, integrating information from images, videos, and social media metadata (Thapa et al., 2025d). This trend is especially relevant in crisis monitoring, misinformation detection, and humanitarian response, where visual and textual signals must be jointly interpreted. At the same time, emerging agentic AI frameworks that combine LLMs with external tools and structured reasoning offer a promising direction for building systems that can autonomously collect, verify, and contextualize event data in dynamic environments (Raheem and Hossain, 2025; Hughes et al., 2025).

In this context, the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE), held at RANLP 2025, continues to serve as a critical platform for advancing interdisciplinary research at the intersection of computational methods and socio-political analysis. Building on the momentum of previous editions (Hürriyetoğlu et al., 2020, 2021b, 2022, 2023, 2024), CASE 2025 highlights innovative approaches for extracting, representing, and interpreting event information spanning traditional NLP pipelines, LLM-centric methods, and multimodal frameworks.

This edition of the workshop features regular

Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025, pages 1–5, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-001

research papers, system descriptions from shared task participants, and keynote talks from experts across disciplines. It also includes a shared task designed to benchmark the capabilities of current systems on complex event extraction problems in multimodal scenarios. This paper provides a brief overview of the CASE 2025 workshop, outlining its themes, activities, and contributions to the broader research community.

2 Accepted Papers

This year, 4 regular papers were accepted. Below, we provide brief descriptions of accepted papers:

- Nadeem et al. (2025) investigate political bias in large language models (LLMs) with a focus on multilingual contexts, particularly across Pakistani languages. Building on the Political Compass Test (PCT), they develop a framework that extracts hidden layer activations from decoder-based models such as Mistral and DeepSeek to identify ideological leanings along economic and social axes. The authors introduce Steering Vector Ensembles (SVE), a representation-level debiasing method that aggregates layer-specific vectors derived from contrastive prompts, enabling inference-time mitigation without fine-tuning. Their experiments show that LLMs encode systematic political bias in internal representations, but SVE effectively reduces this bias, especially in socially framed prompts while preserving fluency and coherence.

- Thapa et al. (2025a) present an extensive survey on socio-political event extraction (SPE), analyzing how advances in natural language processing, machine learning, and LLMs are reshaping the field. The paper systematically reviews datasets, annotation frameworks, extraction methods, and evaluation strategies, highlighting both the progress and persistent challenges in capturing complex, real-world events. The authors emphasize the importance of multilingual and low-resource settings, given the global nature of socio-political events, and point to issues of reproducibility, bias, and ethical concerns in applying SPE systems at scale. They also propose future directions, including leveraging multimodal data, improving temporal and causal reasoning, and aligning event extraction systems with policy and humanitarian needs.

- De Longueville (2025) provides a reflective commentary on the rise of LLMs and their implications for NLP, particularly in the domain of automated socio-political event extraction. They argue that while conversational AI like ChatGPT represents both a revolution and an epiphenomenon for NLP, its significance should be contextualized within decades of technological progress, notably the advent of the Transformer architecture. The paper highlights LLMs' unprecedented zero-shot capabilities and versatility but cautions against overreliance, noting limitations such as high computational cost, hallucinations, sycophancy, and the opacity of LLM-as-a-service deployments. De Longueville (2025) emphasizes that despite the hype, core NLP practices such as precision/recall evaluation, gold-standard datasets, and error analysis remain essential. Ultimately, they conclude that LLMs reshape the landscape of NLP without rendering it obsolete, instead calling for a balanced integration of LLMs with established methodologies and domain-specific knowledge systems.

- Boyd and Mitkov (2025) present a comparative evaluation of rule-based machine translation (RBMT), neural machine translation (NMT), and LLMs for French–English translation using the Europarl corpus. The study employs BLEU and METEOR scores with bootstrap statistical testing, finding NMT, particularly Marian NMT, consistently outperforms LLMs and RBMT in both precision and semantic accuracy, while LLMs trained explicitly for translation (e.g., T5) surpass those with only emergent translation abilities (e.g., LLaMA). RBMT lags far behind in performance, though each approach shows domain-dependent strengths, with NMT best for high-precision needs and LLMs offering versatility for broader, creative applications.

3 Shared Task on Multimodal Content Analysis on Marginalized Sociopolitical Movements

This year's shared task explored multimodal socio-political discourse by focusing on memes, which are an increasingly popular medium for expressing

opinion, humor, and hate online. With the growing role of social media in shaping public perception, this task aims to evaluate systems' abilities to interpret stance, hate, and humor in text-embedded images. The shared task used the PrideMM dataset proposed by Shah et al. (2024). The dataset employs a rigorous annotation schema (Bhandari et al., 2023), consistent with that used in our previous shared tasks (Thapa et al., 2023, 2024) on multimodal content moderation. The task is divided into four subtasks:

- **Subtask A on Hate Speech Detection**: Binary classification of memes as containing Hate Speech or No Hate Speech, using both text and image modalities.
- **Subtask B on Targets of Hate Speech Detection**: Identifies the target of hateful content as Individual, Community, Organization, or Undirected.
- **Subtask C on Topical Stance Detection**: Determines whether the meme Supports, Opposes, or is Neutral toward a marginalized movement.
- **Subtask D on Intended Humor Detection**: Binary classification of memes based on the presence or absence of Intended Humor.

This shared task advances multimodal understanding of contentious online discourse and offers new benchmarks for evaluating models in complex socio-political contexts. Thapa et al. (2025b) provide a detailed overview of the shared task, including participant methods, the task timeline, and a discussion of the key findings. A total of 89 participants took part in the shared task, reflecting strong engagement from the research community. We accepted 13 shared task description papers.

4 Future Direction

As the field of socio-political event extraction continues to evolve, future iterations of the CASE workshop aim to further broaden the scope and impact of the research community. One key direction is the continued exploration of multilingual and multimodal event extraction, recognizing the global and diverse nature of socio-political discourse. Understanding how events manifest across languages and modalities like text, image, video, and audio remains a critical challenge, particularly in crisis monitoring and cross-cultural analysis. We aim to encourage contributions that advance the robustness, adaptability, and inclusivity of event extraction systems in these contexts.

Another important focus will be the integration of alignment techniques in LLM-based event extraction pipelines. As large language models become increasingly central to the field, understanding how to align them with domain-specific ontologies, human feedback, and real-world utility will be crucial. We are particularly interested in prompting strategies, instruction tuning, fine-grained evaluation frameworks, and the use of LLMs within agentic systems that can reason, validate, and act based on extracted event information.

In future editions, we also plan to organize more innovative and task-oriented shared tasks that reflect real-world complexities such as low-resource event extraction, multi-hop event reasoning, and cross-modal fusion. These shared tasks will continue to serve as benchmarks while also driving the development of practical solutions deployable in high-stakes environments. To support and grow the community, we are also looking to introduce mentorship opportunities for early-career researchers and students, especially those from underrepresented regions or working with low-resource languages. We plan to host dedicated mentorship sessions, community-building events, and tutorials to promote inclusion, collaboration, and knowledge transfer across domains.

5 Conclusion

The 8th edition of the CASE workshop highlighted significant progress in socio-political event extraction, with contributions spanning multimodal analysis, large language model applications, and fine-grained stance and hate speech detection. This year's shared task emphasized the growing importance of understanding text-embedded images, reflecting the need to address evolving forms of online discourse. The workshop brought together researchers from diverse disciplines to tackle real-world challenges such as misinformation, polarization, and marginalization through computational methods. Looking ahead, CASE aims to remain an inclusive and interdisciplinary platform that fosters collaboration, supports innovative shared tasks, and promotes research that meaningfully contributes to the understanding of complex socio-political phenomena.

Acknowledgments

Funding for part of this research has been provided by the European Union's Horizon Europe research and innovation programme EFRA [grant number 101093026] and ECO-Ready [grant number 101084201].

Broader Impact

The CASE workshop has a far-reaching impact by promoting interdisciplinary research at the intersection of computational methods and socio-political analysis, encouraging the development of tools and models that can interpret complex societal discourse at scale. By addressing real-world challenges such as hate speech, misinformation, political polarization, and public sentiment, the workshop supports the creation of socially responsible technologies that inform policy, empower marginalized voices, and enhance crisis response. Through shared tasks and diverse participation, CASE promotes equitable access to research opportunities and drives forward the responsible use of NLP for social good.

References

Jacy Reese Anthis, Ryan Liu, Sean M Richardson, Austin C Kozlowski, Bernard Koch, Erik Brynjolfsson, James Evans, and Michael S Bernstein. Position: Llm social simulations are a promising research method. In *Forty-second International Conference on Machine Learning Position Paper Track*.

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 1993–2002.

William Jock Boyd and Ruslan Mitkov. 2025. Machine translation in the ai era: Comparing previous methods of machine translation with large language models. In *Proceedings of The Eighth Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Yan Chen, Kate Sherren, Michael Smit, and Kyung Young Lee. 2023. Using social media images as data in social science research. *New Media & Society*, 25(4):849–871.

Bertrand De Longueville. 2025. Natural language processing vs large language models: this is the end of the world as we know it, and i feel fine. In *Proceedings of The Eighth Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Chenyu Hou, Gaoxia Zhu, Juan Zheng, Lishan Zhang, Xiaoshan Huang, Tianlong Zhong, Shan Li, Hanxiang Du, and Chin Lee Ker. 2024. Prompt-based and fine-tuned gpt models for context-dependent and-independent deductive coding in social annotation. In *Proceedings of the 14th learning analytics and knowledge conference*, pages 518–528.

Yuxin Hou and Junming Huang. 2025. Natural language processing for social science research: A comprehensive review. *Chinese Journal of Sociology*, 11(1):121–157.

Laurie Hughes, Yogesh K Dwivedi, Tegwen Malik, Mazen Shawosh, Mousa Ahmed Albashrawi, Il Jeon, Vincent Dutot, Mandanna Appanderanda, Tom Crick, Rahul De', et al. 2025. Ai agents and agentic systems: A multi-expert analysis. *Journal of Computer Information Systems*, pages 1–29.

Ali Hürriyetoğlu, Osman Mutlu, Erdem Yörük, Farhana Ferdousi Liza, Ritesh Kumar, and Shyam Ratan. 2021a. Multilingual protest news detection - shared task 1, CASE 2021. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 79–91, Online. Association for Computational Linguistics.

Ali Hürriyetoğlu, Hristo Tanev, Osman Mutlu, Surendrabikram Thapa, Fiona Anting Tan, and Erdem Yörük. 2023. Challenges and applications of automated extraction of socio-political events from text (CASE 2023): Workshop and shared task report. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text*, pages 167–175, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Ali Hürriyetoğlu, Hristo Tanev, Vanni Zavarella, Jakub Piskorski, Reyyan Yeniterzi, Osman Mutlu, Deniz Yuret, and Aline Villavicencio. 2021b. Challenges and applications of automated extraction of socio-political events from text (CASE 2021): Workshop and shared task report. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 1–9, Online. Association for Computational Linguistics.

Ali Hürriyetoğlu, Hristo Tanev, Vanni Zavarella, Reyyan Yeniterzi, Osman Mutlu, and Erdem Yörük. 2022. Challenges and applications of automated extraction of socio-political events from text (CASE 2022): Workshop and shared task report. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 217–222, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ali Hürriyetoğlu, Surendrabikram Thapa, Gökçe Uludoğan, Somaiyeh Dehghan, and Hristo Tanev. 2024. A concise report of the 7th workshop on challenges and applications of automated extraction of socio-political events from text. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 248–255.

Ali Hürriyetoğlu, Erdem Yörük, Osman Mutlu, Fırat Duruşan, Çağrı Yoltar, Deniz Yüret, and Burak Gürel. 2021c. Cross-context news corpus for protest event-related knowledge base construction. *Data Intelligence*, 3(2):308–335.

Ali Hürriyetoğlu, Vanni Zavarella, Hristo Tanev, Erdem Yörük, Ali Safaya, and Osman Mutlu. 2020. Automated extraction of socio-political events from news (AESPEN): Workshop and shared task report. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 1–6, Marseille, France. European Language Resources Association (ELRA).

Ariba Khan, Stephen Casper, and Dylan Hadfield-Menell. 2025. Randomness, not representation: The unreliability of evaluating cultural alignment in llms. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 2151–2165.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. 2024. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 6(4):383–392.

Afrozah Nadeem, Mark Dras, and Usman Naseem. 2025. Steering towards fairness: Mitigating political stance bias in llms. In *Proceedings of The Eighth Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Tayiba Raheem and Gahangir Hossain. 2025. Agentic ai systems: Opportunities, challenges, and trustworthiness. In *2025 IEEE International Conference on Electro Information Technology (eIT)*, pages 618–624. IEEE.

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

Hong Shen, Tianshi Li, Toby Jia-Jun Li, Joon Sung Park, and Diyi Yang. 2023. Shaping the emerging norms of using large language models in social computing research. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, pages 569–571.

Xiaoling Shu and Yiwan Ye. 2023. Knowledge discovery: Methods from data mining and machine learning. *Social Science Research*, 110:102817.

Surendrabikram Thapa, Surabhi Adhikari, Hristo Tanev, and Ali Hürriyetoğlu. 2025a. Challenges and applications of automated extraction of socio-political events at the age of large language models. In *Proceedings of The Eighth Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Surendrabikram Thapa, Farhan Jafri, Ali Hürriyetoğlu, Francielle Vargas, Roy Ka-Wei Lee, and Usman Naseem. 2023. Multimodal hate speech event detection-shared task 4, case 2023. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text*, pages 151–159.

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. 2025b. 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)*.

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

Surendrabikram Thapa, Hariram Veeramani, Liang Hu, Qi Zhang, Wei Wang, and Usman Naseem. 2025d. A multimodal prompt-based framework for analyzing code-mixed and low-resource memes. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 1913–1923.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

Challenges and Applications of Automated Extraction of Socio-political Events at the age of Large Language Models

Surendrabikram Thapa[1], Surabhi Adhikari[2], Hristo Tanev[3], Ali Hürriyetoğlu[4]
[1]Virginia Tech, USA, [2]Columbia University, USA,
[3]European Commission, Joint Research Centre, Italy,
[4]Wageningen Food Safety Research, Netherlands
[1]sbt@vt.edu, [2]surabhi.adhikari@columbia.edu,
[3]hristo.tanev@ec.europa.eu, [4]ali.hurriyetoglu@wur.nl

Abstract

Socio-political event extraction (SPE) enables automated identification of critical events such as protests, conflicts, and policy shifts from unstructured text. As a foundational tool for journalism, social science research, and crisis response, SPE plays a key role in understanding complex global dynamics. The emergence of large language models (LLMs) like GPT-4 and LLaMA offers new opportunities for flexible, multilingual, and zero-shot SPE. However, applying LLMs to this domain introduces significant risks, including hallucinated outputs, lack of transparency, geopolitical bias, and potential misuse in surveillance or censorship. This position paper critically examines the promises and pitfalls of LLM-driven SPE, drawing on recent datasets and benchmarks. We argue that SPE is a high-stakes application requiring rigorous ethical scrutiny, interdisciplinary collaboration, and transparent design practices. We propose a research agenda focused on reproducibility, participatory development, and building systems that align with democratic values and the rights of affected communities.

1 Introduction

Socio-political events (SPEs) are occurrences involving political or social actors that have significance for societies or governance. Protests, conflicts, elections, policy changes, and diplomatic interactions are examples of SPEs. In computational terms, an SPE can be represented as a structured record of who did what to whom, when and where, extracted from text (Cai and O'Connor, 2023). Event extraction systems seek to transform unstructured data (e.g. news articles, social media posts) into structured event representations (often as tuples like source–action–target with time and location) (Hu et al., 2024). Such structured event databases enable large-scale analysis of political dynamics and serve as inputs for monitoring conflict, tracking trends, and forecasting crises (Hu et al., 2024). In both academic research and real-world decision-making, having timely and accurate event data is crucial. Analysts use these databases to understand patterns of violence, policymakers use them for early warnings, and humanitarian organizations for situational awareness.

Automated SPE extraction has grown in importance as the volume of text data (news, social media) explodes beyond human coding capacity. Traditional rule-based or supervised systems have been used to populate global event databases (e.g. extracting 'who attacked whom') for decades. Recently, large language models (LLMs) have begun to play a transformative role in this space. LLMs like GPT-3.5 and GPT-4 can, in principle, read and interpret complex texts to identify events with minimal task-specific training. Early experiments show that advanced LLMs (e.g. GPT-4) significantly outperform previous models in zero-shot political event coding, handling nuanced distinctions better and generalizing with fewer examples (Hu et al., 2024). The success of GPT-4 in following event coding guidelines highlights the vast potential of LLMs for this task (Hu et al., 2024). At the same time, LLMs introduce new challenges (like hallucination and transparency issues, discussed later) that must be managed. This position paper takes a hybrid technical and policy-oriented view of automated socio-political event extraction in the era of LLMs, examining not only the algorithmic and data-centric hurdles but also the ethical, legal, and societal implications of these technologies.

2 Technical Challenges in SPE

Despite progress, automated SPE extraction faces numerous technical challenges.

2.1 Ambiguity and Coreference

Language describing socio-political events is often ambiguous. A single phrase can imply different event types depending on context (e.g. "sanction"

Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025, pages 6–19, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-002

could mean an economic sanction or simply approval) (Cai and O'Connor, 2023; Hürriyetoğlu et al., 2022a; Danilova and Popova, 2014). Identifying whether an event actually occurred or is hypothetical (modality) also requires understanding subtle cues (did a politician promise an action or actually do it?). Moreover, the information about one real-world event may be scattered across multiple sentences or reports. Systems must perform coreference resolution to merge mentions referring to the same event. For example, in the text 'A protest broke out in CityX… The demonstration continued into the night', linking 'protest' and 'demonstration' is non-trivial. Recent efforts have been made to explicitly evaluate event coreference linking across sentences (Hürriyetoğlu et al., 2022b). However, ambiguity and cross-sentence reference remain open problems. Without resolving these, an automated system might count one event multiple times or miss it entirely.

2.2 Temporal and Spatial Grounding

Every event entry needs a when and where (Abraham et al., 2018; Westin, 2025). Extracting accurate temporal and geospatial information is challenging. News text may describe an event with relative times ('earlier today', 'last week') that require context (e.g., publication date) to resolve. Locations can be mentioned at various granularities (a city, a region, a country), and many event coders need coordinates, which requires mapping place names to a gazetteer (Hürriyetoğlu et al., 2024). Ensuring that each event is anchored to the correct date and place is vital for analysis (e.g., distinguishing two protests on different days). Temporal ordering (figuring out the sequence of events) is also difficult when texts jump around chronologically. Techniques from temporal IE and geographic entity resolution are needed as part of any robust SPE pipeline. These tasks remain hard, especially in noisy or terse text (like social media), where time/place might not be explicitly stated.

2.3 Multilinguality and Low-Resource Languages

Socio-political events occur worldwide, and being able to extract events from multiple languages is essential for global coverage. Many high-profile event extraction systems have focused on English (or a few major languages) due to data availability. However, relying only on English sources creates a biased pictures (Claro et al., 2019; Miok et al., 2024). The challenge is that NLP resources (annotated data, pretrained models) for many low-resource languages are limited. Progress is being made (Hürriyetoğlu et al., 2022b). Still, performance typically drops for truly low-resource languages (with different scripts or limited data).

2.4 Dataset Quality and Reproducibility

High-quality training and evaluation data are expensive to create (Thapa et al., 2023). Annotating event mentions in text (especially with detailed role labeling or fine-grained event types) is time-consuming and often requires expert knowledge of political contexts (Olsen et al., 2024; Cardie and Wilkerson, 2008). As a result, existing datasets may be small, sparse, or inconsistently annotated. Many academic event extraction datasets (e.g. ACE 2005, TAC KBP event tracks) focus on a limited ontology and are not perfectly aligned with the needs of socio-political analysis (Doddington et al., 2004; Mitamura et al., 2015). On the other hand, political science event datasets (like ICEWS or ACLED) contain high-level coded events but are not released with their source texts (often due to copyright), making it hard to use them for supervised learning or to reproduce results (Raleigh et al., 2010; O'brien, 2010). This raises a reproducibility challenge. A research group may train a model on proprietary news data and output a set of events, but without public text data, others cannot replicate the extraction process. Furthermore, different datasets use different schemas, making it hard to compare systems. Annotation consistency is also an issue, as complex events can suffer from low inter-annotator agreement if guidelines are vague.

2.5 Event Schema and Ontology Design

What counts as an "event" and how it is categorized can vary greatly. Designing an ontology (schema) for events is a foundational challenge that affects extraction (Danilova and Popova, 2014; Xiang and Wang, 2019). SPE extraction has been guided by schemas like CAMEO (Conflict and Mediation Event Observations) which defines a hierarchy of around 20 top-level event classes and over 200 subtypes for political interactions (from cooperative acts like appeals or meetings to conflictual ones like protests, attacks) (Parolin et al., 2019; Gerner et al., 2002). Other ontologies exist (ACE's schema for general events, custom schemas for cybersecurity events, etc.), and social science projects have proposed new ones (e.g., PLOVER, a recent political

violence ontology aligning with CAMEO) (Halterman et al., 2023). The schema design problem has two elements: (1) deciding on the categories and their granularity (balancing detail with annotator reliability), and (2) ensuring models can generalize across schema changes. A rigid ontology may become outdated as new event types emerge (for example, "COVID lockdown protest" might not fit neatly into older categories). On the other hand, very broad definitions reduce analytical usefulness.

3 Applications and Use Cases of LLMs

3.1 Conflict Early Warning and Crisis Forecasting

One of the original motivations for machine-coded event data was to feed conflict early warning systems (Hegre et al., 2019). Projects like the Integrated Crisis Early Warning System (ICEWS) have used continuous streams of coded events (protests, violence, cooperation events, etc.) to predict instability and conflict outbreaks. By analyzing trends e.g. a spike in protests or escalating repressive events, these systems aim to forecast the risk of civil war, mass atrocities, or other crises, enabling preventative action. Automated event extraction greatly speeds up the data pipeline for such systems, which need near-real-time updates from daily news. LLMs could enhance early warning by improving the recall of relevant events (catching subtle precursors in text) and by summarizing situational reports (Foisy et al., 2025; Baek et al., 2023). For example, an LLM might synthesize disparate reports into a narrative of escalating tension.

3.2 Use by Governments and International Organizations

Governments and intergovernmental organizations (IGOs) are heavy users of event data (Ngai et al., 2025). Intelligence and defense agencies use event extraction to monitor global security like identifying terror attacks, troop movements, or diplomatic gestures in open sources. The U.S. government's ICEWS program is one example where automated event data directly supports analysts. Diplomatic services might track protest movements or election-related unrest in real time to inform embassy staff. At the IGO level, organizations like the United Nations or regional bodies (African Union, EU) may utilize event data for peacekeeping and policy decisions (Nohuddin and Zainol, 2020; Amicarelli and Di Salvatore, 2021). The U.N.'s crisis mapping initiatives and the World Bank's political risk assessments rely on understanding the event landscape. Here, comprehensiveness and reliability of event extraction are key. An LLM-powered system might help by reading situation reports or local news in various languages and highlighting events of concern, thus augmenting human analysts.

3.3 NGOs and Humanitarian Monitoring

Non-governmental organizations (NGOs), especially in the human rights and conflict prevention space, have been both producers and consumers of event data (Alhelbawy et al., 2020). A notable example is ACLED (Armed Conflict Location & Event Data Project), an NGO-driven effort that manually curates conflict and protest events across the world. ACLED (Raleigh et al., 2010) and others (e.g. Crisis Group, Human Rights Watch's data teams) might use automated extraction to extend their reach, scanning local media or social platforms for reports of violence that their human coders can then verify and add. Humanitarian organizations can benefit from real-time event feeds to coordinate responses. For instance, knowing about protests turning violent could help the Red Cross prepare, or detecting displacement events could trigger UNHCR action. LLMs could assist these NGOs by quickly summarizing large volumes of community radio transcripts or Facebook posts from affected communities, pulling out events like "village attacked by armed group" or "aid convoy blocked by protesters."

3.4 Event Databases and Knowledge Graphs

In academia and policy research, curated event databases are valuable for studying patterns of conflict, cooperation, and social movements (Zhao et al., 2024). Automated extraction is used to populate and update these databases continuously (Deng et al., 2024; Gottschalk and Demidova, 2018). For example, the GDELT project has attempted to automatically ingest global news and output coded events for every day. While impressive in scale, such efforts sometimes sacrificed precision for breadth. With LLMs, there is potential to improve the quality of automated event databases. An LLM can consider subtler contexts than keyword-based systems, thereby potentially reducing false positives. Moreover, LLMs can help unify or reconcile events. If multiple news reports describe the same protest from different angles, an LLM might consolidate them into one entry with a more com-

plete description (this borders on automatic summarization of events). Knowledge graphs are another use where events can be nodes linking actors, places, and dates in a graph database. Querying such graphs can answer complex questions (e.g. "find all confrontations between government forces and tribe X in the past year"). Automated SPE extraction is what supplies the raw material for these knowledge bases. LLMs could be used to populate new types of relations in graphs, like sentiment or causal links (e.g. "protest led to policy change"). There is active research on using LLMs to enrich knowledge graphs with event information extracted from text (Deng et al., 2024).

3.5 Analytical Tools and Summarization

Finally, a growing application is the use of LLMs for higher-level analysis of event data. Rather than just populating a database, an LLM can help analysts make sense of the data (Kumar et al., 2024). For instance, given a chronology of extracted events, an LLM could produce a narrative report or timeline summary ("In June, a series of protests in X province escalated into clashes by August, prompting government crackdown in September..."). This moves into the realm of report generation and explanatory analysis. Automating such analytical tasks has policy value as busy decision-makers may not have time to read dozens of incident reports, but a well-crafted summary or even an on-demand Q&A powered by an LLM (e.g. "Has violence against civilians increased this month compared to last?") could be immensely helpful. Some prototypes in media monitoring have used LLMs to summarize global news on a topic across countries. For example, summarizing how different countries' press are reacting to a conflict. Those same capabilities can be tuned to summarizing event data. Additionally, interactive exploration via natural language questions is an exciting use case. For example, an analyst could ask the system (which has ingested an event database) questions in English and get answers or charts, without needing to write code or SQL. LLMs can serve as an interface between humans and complex event data, broadening access to insights. Caution is warranted to keep the LLM "grounded" in actual data (so it doesn't fabricate answers). Combining retrieval methods with LLMs (so the model bases answers on retrieved event records) is one technique being explored for this purpose (Arslan et al., 2024).

4 Limitations of LLMs, Multilingual and Global Considerations

4.1 Technical Limitations

Introducing LLMs into the pipeline brings its own set of technical caveats (Thapa et al., 2025). By design, generative LLMs will fill in gaps and produce plausible text even when the input is uncertain. This can lead to hallucinated events, i.e. the model might assert that an event occurred that isn't actually supported by the source (Zhang et al., 2025; Ji et al., 2023; Shiri et al., 2024; Liu et al., 2025). For example, if given a vaguely worded report, an LLM might "assume" a protest happened when in reality the text was speculating. Ensuring faithful extraction requires grounding the LLM to the source text. Relatedly, LLM outputs can be inconsistent; the same prompt might yield slightly different extractions on different runs (due to sampling variability), which is problematic for a deterministic database update. Stability and calibration of confidence in extracted facts are therefore technical issues to solve. Another limitation is interpretability as deep learning models, especially large generative ones, are often black boxes. Understanding why a model classified something as, say, an "attack" versus an "arrest" can be difficult, hindering our ability to trust and refine the system. LLMs also have practical limitations like they may struggle with very long documents (context length limits), or with remembering a long list of ontology definitions without confusion.

4.2 Non-Western Contexts and Local Nuance

Many event extraction tools and models have been developed primarily on Western news sources and in languages like English, Spanish, or French (Aliyu et al., 2024; Kulkarni and Dogra, 2024). Applying these to events in, say, rural Africa or Central Asia can pose problems. The way events are reported, the cultural context, and the actors involved may differ greatly (Hürriyetoğlu et al., 2022b). For example, a "protest" in one country might be described very differently in another country's media (or might not be reported openly at all). Local idioms or euphemisms (e.g., referring to rebel militants as "our boys" in some context) might mask what an event is about. Also, the salience of event types can differ. Events like tribal clashes, land disputes, election violence, etc., each have unique markers. An extraction system needs to be tuned into these nuances. This often requires

involving regional experts in the loop, or at least using region-specific data to fine-tune models. One promising avenue is to engage local journalists or organizations to help create training data (perhaps via annotation or feedback) for their context, creating a more inclusive global system. LLMs, with their ability to absorb vast multi-domain text, might already know some culturally specific references, but careful prompt engineering is needed to make them work for less-covered contexts.

4.3 Cross-Lingual and Low-Resource Techniques

As mentioned, multilingual capability is crucial. There are a few approaches to handle it (Jafri et al., 2024; Alghamdi et al., 2024). One is machine translation (MT), i.e., translate all foreign texts to a pivot language (e.g. English) and then run an English event extractor (Chew et al., 2025; Cabrera, 2024). This was a common strategy in earlier systems, but MT errors can lead to missed or wrong events (especially if translation alters proper names or event verbs). Another approach is using multilingual models like multilingual BERT or XLM (Pires et al., 2019; Conneau et al., 2020), which have some cross-lingual transfer ability. Such models can sometimes be trained on a high-resource language and still be applied to a related low-resource language. Few-shot learning with LLMs could shine where one could prompt an LLM in a target language with a few examples of event annotations in that language (or even in English, relying on its cross-lingual knowledge) and get results. There is early research on prompt-based cross-lingual IE which is encouraging. Additionally, active learning could be employed i.e., the system asks humans to translate or verify a few critical pieces to improve itself iteratively.

4.4 Multimodal Event Extraction

Socio-political events are not only described in text; they may be captured in images, videos, or even satellite data (Bhandari et al., 2023; Thapa et al., 2024). A protest might be live-streamed, a damage assessment might come from satellite imagery, a social media image might show evidence of an attack. Multimodal event extraction seeks to combine text with other data sources to improve event detection and validation. For instance, an automated system could corroborate a reported protest (text) with social media images geotagged in that city showing crowds. LLMs are expanding into multimodal models (e.g. vision-language models like GPT-4's multi-modality or others that can process images) (Thapa et al., 2025; Fei et al., 2024). A future SPE pipeline might take a news article and also any attached photo or video transcript, and use both to decide what happened. Multimodal analysis can improve recall (catch events that text missed but image shows) and precision (disambiguate events by seeing visuals). It also helps in contexts where text might be propagandistic and images can sometimes cut through biases (though they have their own issues of authenticity).

4.5 Bias and Representation in Global Data

Global event extraction must grapple with bias in sources (Xiang and Wang, 2019; Spiliopoulou et al., 2020; Dev et al., 2021). Many regions lack independent media, or any media coverage at all of certain event types (e.g. state repression might be hidden). As a result, automated systems might reflect state narratives or international media agendas. Being aware of these gaps is part of a global perspective. There are efforts to include non-traditional sources. For instance, using reports from NGOs or crowdsourced data to complement news. A balanced approach might merge information from local citizen reports with mainstream media, with the AI model reconciling them. Bias mitigation techniques can be applied, such as calibration (if a known bias exists, adjust the data distribution) (Garrido-Muñoz et al., 2021; Sun et al., 2019). Ultimately, a global system may need regional tuning, as what works well for event extraction in Europe might need rethinking for Central Africa. Community evaluations and workshops (like regional "data challenges") could help identify where current models fall short. Inclusivity in the development process (having NLP researchers and social scientists from diverse regions) is also vital to ensure the tools are attuned to global realities and not just Western media patterns.

5 Policy and Ethical Challenges

5.1 Surveillance and Authoritarian Misuse

A powerful SPE extraction system can turn into a double-edged sword. On one hand, it can provide transparency and early warnings about crises; on the other, it could enable authoritarian surveillance at an unprecedented scale (Yabanci, 2025; Roberts and Oosterom, 2024). Repressive regimes might use automated event detection to track dissident

activities or protests in real-time, flagging leaders and participants for reprisal. Unfortunately, this is not just hypothetical. AI-driven surveillance and policing systems are already used by authoritarian governments and have been found effective in suppressing political unrest and entrenching regimes. If an event extraction tool can scrape social media and news to pinpoint every protest or strike as it begins, authorities could quickly crack down, undermining civil liberties. Even in democratic societies, law enforcement has shown interest in such tools. This kind of proactive surveillance blurs the line between public safety and infringement of the right to assemble.

5.2 Privacy and Human Rights

Related to the above, the privacy implications of large-scale event monitoring are significant (Baldassarre et al., 2024). Socio-political events often involve individuals like protesters, activists, and even victims of violence. If an automated system is parsing social media for events, it might incidentally capture personal data like names of organizers, eyewitness accounts, etc. Even news articles can contain personal identifying information in event descriptions. Using AI to aggregate and analyze this at scale can amplify privacy risks. For instance, extracting a "protest event" from a Facebook post could reveal the poster's political participation without their consent. Furthermore, in conflict zones or authoritarian contexts, being identified in an event report (e.g., as attending a demonstration) could endanger one's safety. Human rights organizations worry that indiscriminate use of such technology could lead to abuses such as compiling watchlists of protesters or surveilling minority communities under the guise of event detection.

5.3 Misinformation and Propaganda

Automated event extraction systems could inadvertently become conduits for misinformation or propaganda if not carefully managed. These systems rely on source data which may be inaccurate or biased. For example, state-controlled media might report a fabricated event (e.g. a false "terror plot foiled") or exaggerate an incident for propaganda. If an automated pipeline naively extracts that into the event database, it lends credence to the false narrative and propagates it to any downstream users (analysts, alert systems, etc.). There is a real risk of false positives where an SPE system could report an event that never actually happened, due to either misinterpretation or malicious input. In the context of political events, such an error can have serious consequences (imagine a system that mistakenly alerts to a "coup attempt" that was just a rumor, and governments could react harshly). Systems should thus cross-validate events with multiple sources or official reports when possible.

5.4 Bias, Fairness, and Data Provenance

Automated SPE extraction inherits and can even amplify biases present in source data (Huang et al., 2024; Kumari et al., 2024). Media reporting bias is well documented. For instance, studies find that international media severely underreport violence in certain regions compared to others. If an event extraction system relies on those media, the resulting database will systematically undercount or underplay conflicts in those underreported regions. This raises fairness concerns around analyses using the data might over-focus on areas that the media highlight and neglect others. Bias can also creep in through the algorithms. If an ML model were trained mostly on, say, Western news text, it might not recognize event triggers in the rhetoric of other cultures or might misclassify events that don't fit its learned patterns. Furthermore, LLMs themselves carry biases from their training data; they might be more likely to extract events that sound "newsworthy" in a Western sense, for example.

6 Recommendations and Guidelines

6.1 Robust Dataset Creation and Sharing

The community should establish best practices for creating and sharing event data. This includes clear documentation of inclusion criteria, coding methodologies, and known limitations of any event dataset. Data collectors (whether researchers or organizations) have a responsibility to explicitly state what sources they use, what counts as an event, and what biases might result. When possible, datasets should be shared in a form that supports reproducibility. For example, reference URLs or source snippets for each coded event (within copyright constraints) should be provided. Creative solutions like releasing machine-readable summaries or embeddings of text can be explored to respect copyright while still enabling method comparison. The community could benefit from an open repository of annotated texts for events (perhaps using texts that are in the public domain or licensed for research) to serve as a benchmark. Moreover, any new event ontology

or schema should ideally be published openly, with rationales for design, to encourage standardization or at least interoperability between projects.

6.2 Integration of LLMs with Human Oversight ("Human-in-the-Loop")

To harness LLM power while safeguarding against errors, a human-in-the-loop approach is highly recommended (Amirizaniani et al., 2024; Cohn et al., 2024). LLMs can be used to draft event annotations or suggest events, but human analysts or annotators should verify critical details, especially for high-impact events. For instance, an LLM might summarize a complex report into a tentative event entry; a human can then check the source, correct any misinterpretation, and approve it. This not only prevents spurious data from entering official records but also allows humans to catch subtle biases the AI might introduce. Output validation is crucial and automated confidence scores from models can guide which events need human review (low confidence or novel event types get flagged). Additionally, employing multiple systems (e.g., an LLM and a rule-based checker) in parallel and comparing outputs where disagreements can be routed to humans can be useful. This kind of cross-validation workflow ensures that LLMs augment rather than replace expert judgment in sensitive applications.

6.3 Transparent Model Use & Explainability

Any use of LLMs or AI for SPE extraction in policy or public-facing contexts should be transparent (Foisy et al., 2025). Stakeholders (from end-users of an event dataset to citizens potentially affected by its use) deserve to know if an event was identified by a human, a classical algorithm, or an LLM, and what the reliability might be. We recommend developing explainability tools specific to event extraction. For example, if an LLM classifies something as an "armed attack" event, the system should ideally provide a rationale or highlight the evidence in text that led to this classification. Techniques such as step-by-step reasoning prompts or modular pipelines can help with interpretability. At the very least, event records generated or assisted by AI could carry a tag or confidence level. In high-stakes use (e.g. legal accountability for conflict incidents), one might decide that no event enters the official record without either two independent sources or human verification similar to journalistic standards. Transparency reports on system performance, biases found, and corrections made would also build trust in the technology.

6.4 Ethical Guidelines and "Do No Harm" Policies

It is imperative to establish and follow ethical guidelines for deploying SPE extraction, particularly in volatile and sensitive regions. Drawing on principles from humanitarian and human rights domains, developers should adopt a "Do No Harm" mentality by anticipating how the technology could cause harm and work to mitigate it. For example, if deploying a system to monitor protests in an oppressive regime, measures should be taken so the data is not easily accessible to the regime to target individuals (perhaps aggregating or anonymizing certain elements). Collaboration with ethics boards or oversight committees can provide external review of such deployments. Access control might be one guideline. For example, sensitive event data (like locations of protest organizers) might only be shared with vetted parties like NGOs, not made fully public. The community could formulate a code of conduct or ethics checklist for SPE projects, including considerations like 'have we accounted for bias?', 'are the communities being monitored aware or have a say?', 'is there a risk of misuse and how are we preventing it?' For LLM-specific issues, guidelines should stress not to over-rely on AI without verification, and to always have a human accountability in the loop for decisions made from event data. When working in conflict zones, respecting local laws and norms, and protecting sources (e.g. journalists or informants who are reporting events) is also part of ethical use.

6.5 Bias Awareness and Correction

To address fairness, we recommend that any large-scale SPE extraction effort include an explicit bias assessment phase. This might involve comparing the AI-extracted data with known baselines (perhaps human-curated datasets like ACLED in some regions) to see where discrepancies lie. If certain event types or areas are consistently under-detected, the model or pipeline should be adjusted (additional training data for those cases, or lowering thresholds). Bias correction techniques such as reweighting events from underrepresented regions can be applied to the output data. Another best practice is involving local stakeholders in evaluating the system's output, like having experts from different regions review the events detected in their region for completeness and accuracy. Not only

does this catch biases, but it also builds a more inclusive system. Data provenance, as mentioned, should be maintained. Each event record ideally links to its source material, which allows users to judge source reliability and bias. If an event comes only from a single source with a strong slant, perhaps the system can flag that (like "source is state media"). Users of the data should be educated on these provenance flags. In essence, continuous auditing for bias and an openness about the system's limits will improve fairness and trustworthiness.

6.6 Collaboration Among Stakeholders

Finally, we urge a strong collaboration between the technical developers (NLP researchers, scientists) and the policy community (political scientists, ethicists, legal experts, and practitioners on the ground). This cross-domain dialogue can ensure that the tools developed address real needs and align with norms. For example, engaging with human rights organizations might highlight the need for certain event categories (like "internet shutdown event") that technologists hadn't considered. Policymakers, on the other hand, should stay informed about the capabilities and limits of the latest tech, avoiding both unrealistic expectations and ungrounded fears. Joint workshops or working groups can produce normative guidelines that marry technical possibilities with ethical guardrails. We recommend formulating clear use policies for different scenarios, e.g., guidelines for using event extraction in election monitoring versus in conflict zones (the latter might require more restraint). By working together on scenario planning, the community can preemptively set standards for responsible use (similar to how bioethics guides biomedical innovations).

7 Future Directions

7.1 Hybrid Extraction Models

Future research will likely explore hybrid models that combine the strengths of LLMs with structured symbolic knowledge (He et al., 2025; Shaik and Doboli, 2025). For example, an LLM could be used to interpret text and draft possible events, but a symbolic reasoner or knowledge graph ensures consistency with known facts (preventing obvious contradictions or impossibilities). Integrating expert-defined rules (from event coding manuals) into LLM prompts or architectures could yield systems that are both flexible and precise. One concrete direction is leveraging existing political ontologies and knowledge bases to guide LLMs, e.g., providing a model with a library of event type definitions and historical examples to reduce ambiguity. This addresses the question posed by researchers like 'can we use expert knowledge to enhance efficiency without extensive new data?'. Progress in prompt engineering and fine-tuning will make LLM outputs more controllable, which is crucial for complex event schemas.

7.2 Adaptive and Continual Learning

Socio-political realities evolve, and so must our extraction systems. A promising avenue is continual learning (Wang et al., 2024) for LLM-based extractors, i.e., the ability to update the model as new event types emerge or new slang/terms enter the lexicon, without forgetting past knowledge. This could involve periodic fine-tuning on newly annotated events or streaming adaptation where the model's prompts are adjusted based on feedback. One challenge is avoiding "catastrophic forgetting" when adapting to new domains (Kirkpatrick et al., 2017). Research into LLMs that can plugin new information (modular learning or using external memory) will benefit SPE greatly, as it means, for example, the system that was never trained on "COVID-19 lockdown protest" could learn that category on the fly. Additionally, ontology evolution should be handled, as event schemas are revised (which happens in social science as new patterns like cyber warfare become relevant), systems need to incorporate those changes.

7.3 Multimodal and Multilingual Fusion

Building on current trends, the future will likely see fully multimodal event extraction in practice. This means models that simultaneously process text, images, video, and maybe audio to detect and validate events. A protest event, for instance, could be confirmed by both a news text and a tweet with a photo. Research into multimodal transformers and alignment techniques (like aligning image detection of violence with text reports) is burgeoning. By 2025 and beyond, we anticipate systems that can, say, take a live social media feed (text + images) and output structured events to dashboards for crisis responders. On the multilingual front, future work may achieve more universal models that work across dozens of languages via a combination of improved training data and leveraging LLM's polyglot capabilities. There is also room for transfer learning between languages and modalities.

For example, an event described in French text and an Arabic tweet might be linked as the same event through a shared embedding space.

7.4 Narrative Construction & Causal Analysis

Moving up the value chain, an exciting research frontier is automated narrative and causality extraction. It's not just about listing events, but understanding how they connect. Future LLM-driven systems could attempt to identify causal or temporal relationships. For example, protest A led to government response B, which triggered conflict C. Some early studies are looking at event chains and temporal reasoning with LLMs. If successful, this could produce draft analytical reports or help populate causal graphs of events, which are immensely useful for political analysis (like understanding escalation paths or conflict dynamics). There is also potential for what-if analysis. With generative models, one could simulate how a sequence of events might unfold under different scenarios, giving policymakers a tool to explore consequences (though this enters speculative territory and would need robust grounding in data). Additionally, as LLMs become more explainable, we might use them to interrogate event data like "Why did violence increase in region X?" and the system might highlight a series of coded events (e.g. arrests, then protests, then clashes) as an explanation. Achieving this level of reliable narrative construction will require advances in discourse understanding and knowledge integration for LLMs.

7.5 Data Responsibility and Ethics

On the policy side, a major future direction is establishing international norms or agreements on the responsible use of AI for social data analysis. Just as there are treaties and agreements on the use of certain surveillance (for instance, UN discussions on digital privacy), we may see efforts to set guidelines for technologies like event extraction, especially as they get more powerful with LLMs. Researchers and practitioners should collaborate in forums to develop a code of ethics specific to computational event monitoring. This could encompass agreements on not facilitating human rights abuses, ensuring data sharing for humanitarian purposes, and perhaps even certification of systems (an independent audit to say an event extraction system meets certain bias and transparency standards). Work in this direction will involve not just technical people, but also lawyers, ethicists, and the communities being monitored. Another aspect is education and literacy. Future efforts should include training for policymakers and journalists on how to interpret AI-generated event data, to avoid misuse or misinterpretation.

7.6 Open Research and Collaboration

Finally, a future direction that underpins all others is maintaining an open and interdisciplinary research environment. The challenges at this sociotechnical junction are complex; solving them will require insights from NLP, machine learning, political science, conflict studies, ethics, and more. We envision more joint research endeavors like political scientists formulating problems that NLP folks can help solve, and NLP advances (like new LLM capabilities) being rapidly tested on social science use cases. There is also likely to be increased benchmarking and evaluation efforts specific to SPE, creating shared tasks that evaluate not just extraction accuracy but also bias, fairness, and utility in downstream analysis. A "roadmap" paper from a multi-disciplinary team could periodically assess where we stand and recalibrate goals (for example, setting a goal to achieve a certain reliability in low-resource languages by year X). As foundation models evolve (e.g., new versions of GPT or open-source LLMs with tens of billions of parameters), continually applying them and assessing their fit for event extraction tasks will be an ongoing process. Keeping this work open (publishing results, sharing models) will ensure broad access and avoid a scenario where only a few large players dominate the technology (which could be risky if their interests don't align with public interest).

8 Conclusion

In conclusion, automated socio-political event extraction sits at a pivotal point with the rise of LLMs. The coming years will likely bring substantial improvements in capability with support for more languages, more nuanced detection, and richer outputs. At the same time, ensuring these advancements are applied responsibly and benefit the global community is a collective task for researchers, practitioners, and policymakers. By recognizing the challenges and actively working on both technical solutions and ethical safeguards, we can harness LLMs to better understand and respond to the socio-political events that shape our world.

References

Susanna Abraham, Stephan Mäs, and Lars Bernard. 2018. Extraction of spatio-temporal data about historical events from text documents. *Transactions in GIS*, 22(3):677–696.

Jawaher Alghamdi, Yuqing Lin, and Suhuai Luo. 2024. Fake news detection in low-resource languages: A novel hybrid summarization approach. *Knowledge-Based Systems*, 296:111884.

Ayman Alhelbawy, Mark Lattimer, Udo Kruschwitz, Chris Fox, and Massimo Poesio. 2020. An nlp-powered human rights monitoring platform. *Expert Systems with Applications*, 153:113365.

Yusuf Aliyu, Aliza Sarlan, Kamaluddeen Usman Danyaro, Abdullahi Sani BA Rahman, and Mujaheed Abdullahi. 2024. Sentiment analysis in low-resource settings: a comprehensive review of approaches, languages, and data sources. *IEEE Access*, 12:66883–66909.

Elio Amicarelli and Jessica Di Salvatore. 2021. Introducing the peacekeeping operations corpus (pkoc). *Journal of Peace Research*, 58(5):1137–1148.

Maryam Amirizaniani, Jihan Yao, Adrian Lavergne, Elizabeth Snell Okada, Aman Chadha, Tanya Roosta, and Chirag Shah. 2024. Developing a framework for auditing large language models using human-in-the-loop. *arXiv preprint arXiv:2402.09346*.

Muhammad Arslan, Saba Munawar, and Christophe Cruz. 2024. Political-rag: using generative ai to extract political information from media content. *Journal of Information Technology & Politics*, pages 1–16.

Edward E Azar. 1980. The conflict and peace data bank (copdab) project. *Journal of Conflict Resolution*, 24(1):143–152.

Seungwon Baek, Do Namgoong, Jinwoo Won, and Seung H Han. 2023. Automated detection of social conflict drivers in civil infrastructure projects using natural language processing. *Applied Sciences*, 13(20):11171.

Maria Teresa Baldassarre, Danilo Caivano, Berenice Fernández Nieto, Domenico Gigante, and Azzurra Ragone. 2024. Fostering human rights in responsible ai: A systematic review for best practices in industry. *IEEE Transactions on Artificial Intelligence*, 6(2):416–431.

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.

Luis Cabrera. 2024. Babel fish democracy? prospects for addressing democratic language barriers through machine translation and interpretation. *American Journal of Political Science*, 68(2):767–782.

Erica Cai and Brendan O'Connor. 2023. A monte carlo language model pipeline for zero-shot sociopolitical event extraction. *arXiv preprint arXiv:2305.15051*.

Claire Cardie and John Wilkerson. 2008. Text annotation for political science research.

Edward Chew, Mahasweta Chakraborti, William Weisman, and Seth Frey. 2025. Machine translation for accessible multi-language text analysis. *Computational Communication Research*, 7(1):1.

Daniela Barreiro Claro, Marlo Souza, Clarissa Castellã Xavier, and Leandro Oliveira. 2019. Multilingual open information extraction: Challenges and opportunities. *Information*, 10(7):228.

Clayton Cohn, Caitlin Snyder, Justin Montenegro, and Gautam Biswas. 2024. Towards a human-in-the-loop llm approach to collaborative discourse analysis. In *International Conference on Artificial Intelligence in Education*, pages 11–19. Springer.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Vera Danilova and Svetlana Popova. 2014. Sociopolitical event extraction using a rule-based approach. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 537–546. Springer.

Songgaojun Deng, Maarten de Rijke, and Yue Ning. 2024. Advances in human event modeling: from graph neural networks to language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6459–6469.

Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, et al. 2021. On measures of biases and harms in nlp. *arXiv preprint arXiv:2108.03362*.

Ling Ding, Xiaojun Chen, Jian Wei, and Yang Xiang. 2023. Mabert: mask-attention-based bert for chinese event extraction. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(7):1–21.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Citeseer.

Hao Fei, Yuan Yao, Zhuosheng Zhang, Fuxiao Liu, Ao Zhang, and Tat-Seng Chua. 2024. From multimodal llm to human-level ai: Modality, instruction, reasoning, efficiency and beyond. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, pages 1–8.

Laurence-Olivier M Foisy, Étienne Proulx, Hubert Cadieux, Jérémy Gilbert, Jozef Rivest, Alexandre Bouillon, and Yannick Dufresne. 2025. Prompting the machine: Introducing an llm data extraction method for social scientists. *Social Science Computer Review*, page 08944393251344865.

Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L Alfonso Ureña-López. 2021. A survey on bias in deep nlp. *Applied Sciences*, 11(7):3184.

Deborah J Gerner, Philip A Schrodt, Omur Yilmaz, and Rajaa Abu-Jabr. 2002. The creation of cameo (conflict and mediation event observations): An event data framework for a post cold war world. In *annual meeting of the American Political Science Association*, volume 29.

Simon Gottschalk and Elena Demidova. 2018. Eventkg: A multilingual event-centric temporal knowledge graph. In *European semantic web conference*, pages 272–287. Springer.

Ralph Grishman and Beth M Sundheim. 1996. Message understanding conference-6: A brief history. In *COLING 1996 volume 1: The 16th international conference on computational linguistics*.

Andrew Halterman, Benjamin E Bagozzi, Andreas Beger, Phil Schrodt, and Grace Scraborough. 2023. Plover and polecat: A new political event ontology and dataset. In *International Studies Association Conference Paper*.

Qiyuan He, Jianfei Yu, and Wenya Wang. 2025. Large language model-enhanced symbolic reasoning for knowledge base completion. *arXiv preprint arXiv:2501.01246*.

Håvard Hegre, Marie Allansson, Matthias Basedau, Michael Colaresi, Mihai Croicu, Hanne Fjelde, Frederick Hoyles, Lisa Hultman, Stina Högbladh, Remco Jansen, et al. 2019. Views: A political violence early-warning system. *Journal of peace research*, 56(2):155–174.

Yibo Hu, Erick Skorupa Parolin, Latifur Khan, Patrick Brandt, Javier Osorio, and Vito D'Orazio. 2024. Leveraging codebook knowledge with nli and chatgpt for zero-shot political relation classification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 583–603.

Nannan Huang, Haytham Fayek, and Xiuzhen Zhang. 2024. Bias in opinion summarisation from pretraining to adaptation: A case study in political bias. *arXiv preprint arXiv:2402.00322*.

Ali Hürriyetoğlu, Osman Mutlu, Fatih Beyhan, Fırat Duruşan, Ali Safaya, Reyyan Yeniterzi, and Erdem Yörük. 2022a. Event coreference resolution for contentious politics events. *arXiv preprint arXiv:2203.10123*.

Ali Hürriyetoğlu, Osman Mutlu, Fırat Duruşan, Onur Uca, Alaeddin Gürel, Benjamin J. Radford, Yaoyao Dai, Hansi Hettiarachchi, Niklas Stoehr, Tadashi Nomoto, Milena Slavcheva, Francielle Vargas, Aaqib Javid, Fatih Beyhan, and Erdem Yörük. 2022b. Extended multilingual protest news detection - shared task 1, CASE 2021 and 2022. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 223–228, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ali Hürriyetoğlu, Surendrabikram Thapa, Gökçe Uludoğan, Somaiyeh Dehghan, and Hristo Tanev. 2024. A concise report of the 7th workshop on challenges and applications of automated extraction of socio-political events from text. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 248–255.

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*.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Amit Kulkarni and Varun Dogra. 2024. Comprehensive survey of event extraction methods in natural language processing. In *2024 International Conference on Sustainable Communication Networks and Application (ICSCNA)*, pages 925–929. IEEE.

Raghvendra Kumar, Ritika Sinha, Sriparna Saha, and Adam Jatowt. 2024. Extracting the full story: a multimodal approach and dataset to crisis summarization in tweets. *IEEE Transactions on Computational Social Systems*.

Gitanjali Kumari, Anubhav Sinha, Asif Ekbal, Arindam Chatterjee, and Vinutha B N. 2024. Enhancing the fairness of offensive memes detection models by mitigating unintended political bias. *Journal of Intelligent Information Systems*, 62(3):735–763.

Lincan Li, Jiaqi Li, Catherine Chen, Fred Gui, Hongjia Yang, Chenxiao Yu, Zhengguang Wang, Jianing Cai, Junlong Aaron Zhou, Bolin Shen, et al. 2024. Political-llm: Large language models in political science. *arXiv preprint arXiv:2412.06864*.

Jiangwei Liu, Liangyu Min, and Xiaohong Huang. 2021. An overview of event extraction and its applications. *arXiv preprint arXiv:2111.03212*.

Wenxuan Liu, Zixuan Li, Long Bai, Yuxin Zuo, Daozhu Xu, Xiaolong Jin, Jiafeng Guo, and Xueqi Cheng. 2025. Towards event extraction with massive types: Llm-based collaborative annotation and partitioning extraction. *arXiv preprint arXiv:2503.02628*.

Mingyu Derek Ma, Jiao Sun, Mu Yang, Kung-Hsiang Huang, Nuan Wen, Shikhar Singh, Rujun Han, and Nanyun Peng. 2021. Eventplus: A temporal event understanding pipeline. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 56–65.

Charles McClelland. 1978. World event/interaction survey, 1966-1978. *WEIS Codebook ICPSR*, 5211(640):49.

Kristian Miok, Encarnación Hidalgo Tenorio, Petya Osenova, Miguel-Angel Benitez-Castro, and Marko Robnik-Šikonja. 2024. Multi-aspect multilingual and cross-lingual parliamentary speech analysis. *Intelligent Data Analysis*, 28(1):239–260.

Teruko Mitamura, Zhengzhong Liu, and Eduard H Hovy. 2015. Overview of tac kbp 2015 event nugget track. In *TAC*.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Eric WT Ngai, Ariel KH Lui, and Brian CW Kei. 2025. Natural language processing in government applications: a literature review and a case analysis. *Industrial Management & Data Systems*, 125(6):2067–2104.

Puteri N.E. Nohuddin and Zuraini Zainol. 2020. Discovering explicit knowledge using text mining techniques for peacekeeping documents. *Int. J. Bus. Inf. Syst.*, 35(2):152–166.

Sean P O'brien. 2010. Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International studies review*, 12(1):87–104.

Helene Olsen, Étienne Simon, Erik Velldal, and Lilja Øvrelid. 2024. Socio-political events of conflict and unrest: A survey of available datasets. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 40–53.

Erick Skorupa Parolin, Sayeed Salam, Latifur Khan, Patrick Brandt, and Jennifer Holmes. 2019. Automated verbal-pattern extraction from political news articles using cameo event coding ontology. In *2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE intl conference on high performance and smart computing,(HPSC) and IEEE intl conference on intelligent data and security (IDS)*, pages 258–266. IEEE.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Clionadh Raleigh, Rew Linke, Håvard Hegre, and Joakim Karlsen. 2010. Introducing acled: An armed conflict location and event dataset. *Journal of peace research*, 47(5):651–660.

Tony Roberts and Marjoke Oosterom. 2024. Digital authoritarianism: a systematic literature review. *Information Technology for Development*, pages 1–25.

Philip A Schrodt. 2001. Automated coding of international event data using sparse parsing techniques. In *annual meeting of the International Studies Association, Chicago*.

Philip A Schrodt, Shannon G Davis, and Judith L Weddle. 1994. Political science: Keds—a program for the machine coding of event data. *Social Science Computer Review*, 12(4):561–587.

Hashmath Shaik and Alex Doboli. 2025. Using a symbolic knowledge graph to address llm limitations in analog circuit topology generation. In *2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 00528–00533. IEEE.

Fatemeh Shiri, Farhad Moghimifar, Reza Haffari, Yuan-Fang Li, Van Nguyen, and John Yoo. 2024. Decompose, enrich, and extract! schema-aware event extraction using llms. In *2024 27th International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE.

Evangelia Spiliopoulou, Salvador Medina Maza, Eduard Hovy, and Alexander Hauptmann. 2020. Event-related bias removal for real-time disaster events. *arXiv preprint arXiv:2011.00681*.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. From humans to machines: can chatgpt-like llms effectively replace human annotators in nlp tasks. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*. Association for the Advancement of Artificial Intelligence.

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

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2024. A comprehensive survey of continual learning: Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5362–5383.

Fereshta Westin. 2025. Time, technique and text: scoping review of temporal information extraction and categorisation in documents. *Journal of Documentation*, 81(7):135–156.

Wei Xiang and Bang Wang. 2019. A survey of event extraction from text. *IEEE Access*, 7:173111–173137.

Bilge Yabanci. 2025. Surveil, datafy, publicize: digital authoritarianism and migration governance in turkey. *Democratization*, 32(4):1016–1041.

Ziyao Zhang, Chong Wang, Yanlin Wang, Ensheng Shi, Yuchi Ma, Wanjun Zhong, Jiachi Chen, Mingzhi Mao, and Zibin Zheng. 2025. Llm hallucinations in practical code generation: Phenomena, mechanism, and mitigation. *Proceedings of the ACM on Software Engineering*, 2(ISSTA):481–503.

Bang Zhao, Yilong Zhao, and Ying Mao. 2024. A method for judicial case knowledge graph construction based on event extraction. In *Proceedings of the 2024 9th International Conference on Intelligent Information Technology*, pages 62–69.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Appendix

A.1 Related Works

Systematic political event data collection dates back to the Cold War era. In the 1960s and 70s, political scientists began manual coding of international events from news reports to enable quantitative analysis (Olsen et al., 2024). Influential early datasets like WEIS (World Events Interaction Survey) and COPDAB (Conflict and Peace Data Bank) catalogued interstate events (e.g. protests, conflicts, diplomatic acts) by human annotation of news archives (McClelland, 1978; Olsen et al., 2024; Azar, 1980). These pioneering efforts demonstrated the value of structured event data but were labor-intensive and limited in scope (covering only certain actors or regions). By the late 1980s, researchers recognized that much of this coding could be automated by text processing. The Kansas Event Data System (KEDS) in the early 1990s was a seminal rule-based system that used dictionaries and patterns to code events from newswire feeds (like Reuters) (Schrodt et al., 1994). KEDS (and its successor TABARI) could scan sentences for keywords indicating actions (e.g. 'attack', 'meet') and map them to predefined event types, initiating the era of machine-coded event databases (Schrodt, 2001). These early systems were capable of coding thousands of articles, paralleling developments in the NLP field of information extraction.

In the 1990s and 2000s, the NLP community's work on event extraction evolved in parallel. Early information extraction (IE) tasks in NLP, such as the MUC competitions and later ACE, involved identifying event "triggers" and participants in text (for example, extracting a terrorist bombing event with its perpetrator, target, date, etc.) (Grishman and Sundheim, 1996; Doddington et al., 2004). While political scientists' event databases aimed at capturing abstract real-world events (often aggregating information across sources), NLP tasks focused on text-bound events with token-level annotations (Olsen et al., 2024). This led to a divergence. Socio-political event databases prioritized what actually happened in the world (even if details were spread across multiple documents), whereas NLP event annotations captured what was explicitly mentioned in a single text. Nonetheless, by the 2010s there was convergence in methodology. Statistical and ML-based approaches emerged for event extraction. For example, supervised classifiers to detect if a sentence describes a protest,

or sequence labeling models to mark event triggers and arguments. Researchers began applying emerging deep learning techniques to event extraction, achieving improvements over brittle pattern-matchers (Olsen et al., 2024). However, these supervised models required substantial annotated data (which was scarce for fine-grained socio-political events) and often struggled to adapt when event schemas or ontologies changed.

The late 2010s and early 2020s saw the advent of large pretrained language models, culminating in today's LLMs (Thapa et al., 2025; Naveed et al., 2023). Initially, these models were used as contextual encoders in neural event extraction pipelines (Hu et al., 2024; Ma et al., 2021; Ding et al., 2023). For example, BERT-based classifiers for protest detection or relational models for 'who did what to whom' (Liu et al., 2021). More recently, prompt-based extraction and in-context learning have become feasible. Given a prompt describing event categories or a few examples, an LLM can attempt to parse new texts into structured event records without explicit retraining. This zero-shot or few-shot capacity is attractive for socio-political events, which often require flexibility to new event types or languages. Early studies are mixed but promising. For instance, one study found GPT-4 could achieve nearly the performance of a supervised classifier in coding political event types, and even exceeded some rule-based systems in recall (Hu et al., 2024). At the same time, prompting LLMs for complex, fine-grained event coding exposes issues (memory limits for long ontology descriptions, prompt sensitivity, etc.), indicating that LLMs are not a silver bullet (Thapa et al., 2025; Li et al., 2024; Ziems et al., 2024). The field has now reached a point where hybrid approaches are being explored like combining LLMs with knowledge bases, using retrieval-augmented generation (RAG) for factual grounding, and integrating human feedback for higher fidelity. This sets the stage for understanding the technical challenges that persist and the new considerations that arise in the LLM era.

Multimodal Hate, Humor, and Stance Event Detection in Marginalized Sociopolitical Movements

Surendrabikram Thapa[1], Siddhant Bikram Shah[2], Kritesh Rauniyar[3, 4], Shuvam Shiwakoti[1], Surabhi Adhikari[5], Hariram Veeramani[6], Kristina T. Johnson[2], Ali Hürriyetoğlu[7], Hristo Tanev[8], Usman Naseem[9]

[1]Virginia Tech, USA, [2]Northeastern University, USA,
[3]Delhi Technological University, India, [4]IIMS College, Nepal, [5]Columbia University, USA,
[6]UCLA, USA, [7]Wageningen Food Safety Research, Netherlands,
[8]European Commission, Joint Research Centre, Italy, [9]Macquarie University, Australia
[1]{surendrabikram, shuvam}@vt.edu, [3]rauniyark11@gmail.com,
[7]ali.hurriyetoglu@wur.nl, [8]hristo.tanev@ec.europa.eu

Abstract

This paper presents the Shared Task on Multimodal Detection of Hate Speech, Humor, and Stance in Marginalized Socio-Political Movement Discourse, hosted at CASE 2025. The task is built on the PrideMM dataset, a curated collection of 5,063 text-embedded images related to the LGBTQ+ pride movement, annotated for four interrelated subtasks: (A) Hate Speech Detection, (B) Hate Target Classification, (C) Topical Stance Classification, and (D) Intended Humor Detection. Eighty-nine teams registered, with competitive submissions across all subtasks. The results show that multimodal approaches consistently outperform unimodal baselines, particularly for hate speech detection, while fine-grained tasks such as target identification and stance classification remain challenging due to label imbalance, multimodal ambiguity, and implicit or culturally specific content. CLIP-based models and parameter-efficient fusion architectures achieved strong performance, showing promising directions for low-resource and efficient multimodal systems.

1 Introduction

In the ever-evolving digital landscape, social media has become a pivotal arena for discourse, particularly for marginalized socio-political movements (Bhandari et al., 2023; Shiwakoti et al., 2024). Within these online spaces, text-embedded images, like memes, have emerged as a powerful and prevalent medium of communication. They serve as potent vehicles for expressing solidarity, fostering resistance, and shaping attitudes and perceptions both within and beyond these communities. The multimodal nature of memes, combining imagery and text, presents a formidable challenge for machine learning systems, which must move beyond simplistic analyses to grasp the multifaceted expressions conveyed (Pramanick et al., 2021b). As platforms increasingly grapple with content moderation challenges, the ambiguity between satire and offense in such imagery underscores a critical gap in computational analysis: multimodal understanding must disentangle layered communicative intents to mitigate harm while preserving cultural context (Scott, 2021).

The discourse surrounding marginalized communities is often complex and multifaceted, where the lines between humor, satire, and genuine harm are frequently blurred (Klassen and Fiesler, 2022). Memes, in this context, can simultaneously be instruments of empowerment and weapons of oppression, making the task of content moderation exceptionally difficult. A single label often fails to capture the layered meanings embedded within these images. Consequently, there is a pressing need for a more nuanced, multi-aspect understanding of such content to develop more effective AI systems (Pramanick et al., 2021a). The PrideMM dataset epitomizes this complexity, centering on discourse surrounding the LGBTQ+ movement where memes frequently blur the lines between humor and harm (Shah et al., 2024).

To address this critical research gap, building on our previous shared tasks at CASE 2024 (Thapa et al., 2024b; Hürriyetoğlu et al., 2024) and CASE 2023 (Thapa et al., 2023a; Hürriyetoğlu et al., 2023) we present the Shared Task on Multimodal Detection of Hate Speech, Humor, and Stance in Marginalized Socio-Political Movement Discourse at CASE 2025 (Hürriyetoğlu et al., 2025). This task utilizes the PrideMM dataset, a curated collection of memes related to the LGBTQ+ pride movement. The shared task is designed to spur the development of models that can analyze text-embedded images from four distinct yet interconnected perspectives: 1) Detection of Hate Speech, 2) Classifying the

Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025, pages 20–31, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-003

Targets of Hate Speech, 3) Classification of Topical Stance, and 4) Detection of Intended Humor. We frame these subtasks together to encourage holistic approaches that can capture the entangled social, cultural, and affective dimensions of online content.

This paper provides a comprehensive overview of the shared task, including a description of the dataset, the evaluation metrics for each subtask, a summary of the participating teams and their methodologies, and an analysis of the results. Through this shared task, we aim to foster innovation in multimodal analysis and contribute to the development of more sophisticated and context-aware models for understanding online discourse.

2 Dataset

For our shared task, we utilize the PrideMM dataset, as shown in Table 1, introduced by Shah et al. (2024), which comprises a total of 5,063 text-embedded images (memes, posters, and infographics) related to the LGBTQ+ Pride movement. This multimodal dataset addresses the need for more inclusive and nuanced resources in the meme analysis space by encompassing four distinct yet related tasks: (A) Hate Speech Detection, (B) Hate Target Classification, (C) Topical Stance Classification, and (D) Humor Detection. The dataset spans content from 2020 to 2024, collected from Twitter, Facebook, and Reddit using targeted queries and community-specific groups. To ensure high-quality and diverse samples, redundant images were filtered using deduplication tools, and OCR was applied to extract and clean embedded text. Each image in the dataset was independently annotated by five trained annotators through a rigorous three-phase annotation protocol to enhance label consistency. The annotators labeled every image for hate presence, stance, and humor, and for hateful images, also annotated the target (undirected, individual, community, or organization).

3 Shared Task Description

Social media platforms amplify controversial content and quickly disseminate conflict. Meanwhile, humorous content, especially memes, has become more popular as a tool for digital community involvement and influence (Pramanick et al., 2021a). As people become more outspoken about their opinions, stance identification is essential when analyzing public opinion (AlDayel and Magdy, 2021).

Subtask	Classes	Train	Eval	Test	Total
Subtask A	Hate	1,985	248	249	2,482
	No-Hate	2,065	258	258	2,581
Subtask B	Undirected	617	77	77	771
	Individual	199	25	25	249
	Community	931	116	117	1,164
	Organization	238	30	30	298
Subtask C	Neutral	1,166	146	146	1,458
	Support	1,527	191	191	1,909
	Oppose	1,357	169	170	1,696
Subtask D	No Humor	1,313	164	165	1,642
	Humor	2,737	342	342	3,421

Table 1: Statistics of the dataset provided to the participants as part of the shared task.

This shared task focuses on investigating the potential of different multimodal models and the identification of four unique tasks related to socio-political discourse. Further details on subtasks can be found below:

3.1 Subtask A: Hate Speech Detection

The primary objective of this subtask is to identify the presence of hate speech in the text-embedded images. This is a binary classification task, where each sample is annotated with one of two possible labels: *Hate Speech* or *No Hate Speech*. The dataset used for this task primarily concentrates on visuals in which text is important to expressing meaning, facilitating a more nuanced analysis of harmful or offensive content. Using both textual and visual, the dataset offers clear separation between the two categories.

3.2 Subtask B: Targets of Hate Speech Detection

For content identified as hateful, this subtask requires a more granular analysis to determine the target of the hate speech in the images. The image contains hateful text, with the determined target categories as *Individual*, *Community*, *Organization*, or *Undirected*. This subtask focuses on four specific categories within the text-embedded images, which help to identify and understand the type of hateful content.

3.3 Subtask C: Topical Stance Detection

This subtask aims to classify the stance of the meme towards the marginalized movement itself, with possible labels of *Support*, *Oppose*, or *Neutral*. This subtask involves classifying and understanding the stance of the meme images, with a focus on understanding the type of stance that facilitates

grasping the categories of stance.

3.4 Subtask D: Intended Humor Detection

Recognizing the prevalence of satire and humor in this domain, this subtask challenges participants to identify whether a meme is intended to be humorous, using *Humor* or *No Humor* labels.

4 Participants' Methods

4.1 Overview

Of the 89 registered participants, 21 submitted results for Subtask A, 14 for Subtask B, 13 for Subtask C, and 16 for Subtask D. The leaderboards for these subtasks are presented in Table 2, table 3, table 4 and table 5.

4.2 Methods

The following section presents brief overviews of the participating teams' approaches, based on the methodologies outlined in their system description papers.

4.2.1 Subtask A: Hate Speech Detection

TSR (Ray et al., 2025) presented FIMIF (Feature Interaction for Multi-Modal Integration and Fusion), a lightweight multimodal framework that relies on frozen CLIP ViT-L/14 encoders for extracting text and image embeddings. These embeddings were compressed into low-dimensional spaces using residual projection layers before being passed to a multiplicative feature interaction module designed to capture higher-order cross-modal relationships. Their approach emphasizes efficiency, with only 25k–51k trainable parameters, yet it achieved an F1-score of 81.85% and accuracy of 81.85% in hate speech detection. This demonstrates that dimensionality compression coupled with multiplicative fusion can yield competitive results on multimodal hate classification.

PhantomTroupe (Amin et al., 2025) experimented with multiple approach, including unimodal and multimodal, where the fine-tuned Qwen2.5-VL-7B-Instruct- bnb-4bit using the unsloth framework outperformed acheiving the F1-score of 80.86%. Both approaches followed a transformer-based model, placing the team in 5th position.

MemeMasters (Shakya and Gurung, 2025) utilized a fine-tuned CLIP model as their primary architecture. Their approach involved concatenating the visual and textual embeddings from the CLIP model and feeding them into a lightweight classification head. Using their standard configuration without task-specific modifications, the system achieved a macro F1-score of 80%, showing consistent and balanced performance across both the "Hate" and "Non-Hate" classes.

Multimodal Kathmandu (Maharjan et al., 2025) employed a Co-Attention Ensemble architecture built upon frozen CLIP-ViT features. Text and image embeddings were concatenated and passed through multi-layer Transformer encoders, with predictions averaged across five ensemble members to reduce variance. This approach achieved an F1-score of 79.29% and an accuracy of 79.29%, highlighting the robustness of variance-reduction strategies for multimodal hate speech detection.

MLInitiative (Acharya et al., 2025) investigated two multimodal architectures, a ResNet-18 with BERT model and a SigLIP2 model, for hate speech detection. Their fine-tuned SigLIP2 model outperformed the ResNet-18+BERT baseline, achieving an F1-score of 79.27%. This performance placed their system 9th on the final leaderboard for the subtask.

ID4Fusion (Rashfi et al., 2025) utilized transformer-based models, RoBERTa and HateBERT were fine-tuned for text analysis, while EfficientNet-B7 and Vision Transformer (ViT) were utilized for images. The predictions from these models were integrated using a late-fusion ensemble approach, providing more weight to textual features compared to visual features. In the leaderboard, they secured 10th position.

Silver (Mainali et al., 2025) evaluated a range of unimodal and multimodal models, including transformer-based text models like BERT and ROBERTa, CNN-based vision models like DenseNet and EfficientNet, and fusion methods like CLIP. Their results demonstrated that multimodal systems performed better than unimodal baselines, with a CLIP-based model achieving the top macro F1-score of 78.28%. The authors noted that models often misclassified sarcastic or ironic content where hate was conveyed visually rather than through explicit text.

Rank	Team Name	Codalab Username	F1-score (%)	Accuracy (%)	Precision (%)	Recall (%)
1	TJU-MI	wangxiuxian	84.22	84.22	84.22	84.22
2	-	Ryuan	82.84	82.84	82.91	82.88
3	IMU-L	jiaranDiana	82.05	82.05	82.17	82.11
4	TSR (Ray et al., 2025)	ray-sushant	81.85	81.85	81.91	81.89
5	Phantom Troupe (Amin et al., 2025)	Neuron-Force	80.86	80.87	80.86	80.86
6	MemeMasters (Shakya and Gurung, 2025)	shrutigurung	80.05	80.08	80.12	80.04
7	Multimodal Kathmandu (Maharjan et al., 2025)	Sujal_Maharjan	79.29	79.29	79.33	79.32
8	-	NextTry	79.28	79.29	79.28	79.28
9	MLInitiative (Acharya et al., 2025)	ankitbk07	79.27	79.29	79.30	79.27
10	ID4Fusion (Rashfi et al., 2025)	Rashfi	78.68	78.70	78.70	78.68
11	Silver (Mainali et al., 2025)	rohanmainali	78.28	78.30	78.33	78.27
12	MMFusion (Rane, 2025)	prerana3	77.89	77.91	78.14	77.99
13	CUET NOOB (Joy et al., 2025)	TomalJoy	74.16	74.16	74.16	74.17
14	-	Tanvir_77	74.02	74.16	74.47	74.05
15	Overfitters (Bhattarai et al., 2025)	bidhancb	73.77	73.77	73.82	73.80
16	-	AkshYat	73.37	73.37	73.47	73.42
17	Luminaries (Esackimuthu, 2025)	akshayy22	72.17	72.19	72.34	72.25
18	YS	ysb	69.23	69.23	69.27	69.26
19	MLP (Verma and Kumar, 2025)	Durgeshverma24iitram	66.02	66.27	66.54	66.14
20	wangkongqiang (Kongqiang and Peng, 2025)	wangkongqiang	62.09	63.31	65.91	63.65
21	Musafir	MDSagorChowdhury	58.28	62.33	68.62	61.79

Table 2: Sub-task A (Hate Speech Classification) Leaderboard, Ranked by Macro F1-score. All scores are presented as percentages (%). It is to be noted that this leaderboard contains the score till the test deadline and does not consider further runs done by participants as a part of the system description paper.

MMFusion (Rane, 2025) implemented a multimodal architecture using RoBERTa-base for textual analysis and a ResNet50 model for visual feature extraction. These features were projected into a shared dimensional space and combined using an 8-head multi-head attention mechanism to capture cross-modal interactions. The team also employed focal loss to concentrate on difficult samples and used a test-time augmentation (TTA) strategy to improve robustness, ultimately achieving an F1-score of 77.8%.

CUET NOOB (Joy et al., 2025) used the multimodal attention-based late fusion approach to capture cross-modal interactions. The model achieved an F1-score of 74.16%, ranking 13th overall. The authors also experimented with unimodal models like DistilBERT for text and ViT for images.

Overfitters (Bhattarai et al., 2025) utilized a multimodal fusion model named BERTRES for hate speech detection, combining textual features from a BERT-base model with visual embeddings from a ResNet-50 model. The concatenated feature vector was processed through a classifier with separate heads for each task. This approach achieved an F1-score of 73.77%, placing them 15th in the task. The paper suggests the model's performance was limited by the difficulty of capturing subtle, implied hate and sarcasm in memes.

Luminaries (Esackimuthu, 2025) explored a hybrid modeling approach by combining the ALBERT-base v2 transformer with classical machine learning models such as XGBoost, LightGBM, Gradient Boosting, and MLP classifiers. Predictions from these systems, trained on TF-IDF and syntactic features alongside contextual embeddings, were integrated through a weighted ensembling strategy. This ensemble achieved an F1-score of 72.17%, ranking 17th overall. The authors note that ensembling effectively leveraged complementary strengths across models, though additional linguistic features and further tuning of ensemble weights could yield improvements.

MLP (Verma and Kumar, 2025) developed multimodal frameworks that fused XLM-RoBERTa and ViT embeddings with attention-based fusion, as well as alternative combinations with CLIP and BERT encoders. Their best-performing configuration achieved an F1-score of 66.02% and an accuracy of 66.27%. The system demonstrated the effectiveness of early fusion and cross-modal attention in detecting hate content from memes.

wangkongqiang (Kongqiang and Peng, 2025) employed different approaches, including an ensemble model integrating text and image features (utilizing BERT, XLNet, and InceptionNet), a K-max pooling neural network utilizing pre-trained GloVe embeddings and cyclic learning rate scheduling, and a multinomial naive Bayes (MNB) model. The MNB achieved an F1-score of 62.09%, which placed them in 20th position.

Rank	Team Name	Codalab Username	F1-score (%)	Accuracy (%)	Precision (%)	Recall (%)
1	TJU-MI	wangxiuxian	65.30	64.26	67.59	63.83
2	-	Ryuan	63.35	64.26	65.56	62.04
3	TSR (Ray et al., 2025)	ray-sushant	60.57	63.05	61.69	60.21
4	IMU-L	jiaranDiana	60.15	63.05	62.30	60.38
5	Multimodal Kathmandu (Maharjan et al., 2025)	Sujal_Maharjan	57.77	58.23	56.66	59.22
6	Overfitters (Bhattarai et al., 2025)	bidhancb	56.28	57.03	54.07	60.32
7	MMFusion (Rane, 2025)	prerana3	55.39	59.04	56.53	55.04
8	MLInitiative (Acharya et al., 2025)	ankitbk07	54.86	58.23	60.44	52.49
9	MemeMasters (Shakya and Gurung, 2025)	shrutigurung	51.50	53.82	54.27	50.59
10	Silver (Mainali et al., 2025)	rohanmainali	50.18	51.81	50.92	54.22
11	Luminaries (Esackimuthu, 2025)	akshayy22	49.84	55.42	52.89	48.69
12	Musafir	MDSagorChowdhury	37.93	44.18	40.08	41.43
13	wangkongqiang (Kongqiang and Peng, 2025)	wangkongqiang	34.53	47.79	55.52	33.22
14	MLP (Verma and Kumar, 2025)	Durgeshverma24iitram	27.39	40.96	31.58	27.57

Table 3: Sub-task B (Target Identification for Hate Speech) Leaderboard, Ranked by Macro F1-score. All scores are presented as percentages (%). It is to be noted that this leaderboard contains the score till the test deadline and does not consider further runs done by participants as a part of the system description paper.

4.2.2 Subtask B: Targets of Hate Speech Detection

TSR (Ray et al., 2025) adapted their FIMIF pipeline by addressing the severe class imbalance in the dataset. They incorporated weighted cross-entropy loss and deterministic oversampling of minority classes to stabilize learning across the four target categories (Undirected, Individual, Community, Organization). The system combined CLIP-based representations through residual and multiplicative modules, reaching an F1-score of 60.57% and accuracy of 63.05%. The results indicate that their compact architecture could model nuanced target categories effectively despite relying on a parameter-efficient design.

Multimodal Kathmandu (Maharjan et al., 2025) designed a Hierarchical Cross-Attention Transformer that allowed textual tokens to query visual regions directly. This task-specific architecture reached an F1-score of 57.77% with an accuracy of 58.23%, ranking 5th on the leaderboard. The results underscore the value of explicit cross-modal grounding for distinguishing between Community, Individual, Organization, and Undirected targets.

Overfitters (Bhattarai et al., 2025) applied their BERTRES architecture, which fuses features from BERT and ResNet-50. A key part of their methodology was the use of a class-weighted cross-entropy loss to mitigate the skewed label distribution present in the dataset for this subtask. This strategy contributed to their strongest performance, securing the 6th position with an F1-score of 56.28%.

MMFusion (Rane, 2025) utilized their RoBERTa-ResNet50 architecture with cross-modal attention for the target identification task. To address the significant class imbalance in this subtask, they used a focal loss function with class-specific weighting in addition to a test-time augmentation strategy. The model struggled to differentiate between certain categories, particularly confusing "Individual" targets with "Community" targets, and achieved a final F1-score of 55.3%.

MLInitiative (Acharya et al., 2025) applied two multimodal systems: a combined ResNet-18 and BERT architecture and a SigLIP2 model. The SigLIP2 model proved to be superior, securing an F1-score of 54.86% and ranking 8th on the leaderboard. The authors note that both models performed relatively poorly on this task, attributing the difficulty to the imbalanced nature of the associated dataset.

MemeMasters (Shakya and Gurung, 2025) adapted their fine-tuned CLIP model for the target classification task by applying over-sampling to mitigate the severe class imbalance present in the dataset. The model struggled with the fine-grained nature of this task, showing the lowest recall for the "Individual" class and frequent confusion between the "Undirected" and "Community" targets. This resulted in a macro F1-score of 52%.

Silver (Mainali et al., 2025) addressed the target classification task, noting it was particularly challenging due to the highly uneven distribution of classes and the subjective nature of defining a target. Comparing various unimodal and multimodal systems, their CLIP-based model again achieved the best performance with a macro F1-score of 56.30%. This score represented a considerable performance decline compared to other subtasks, with the paper highlighting

Rank	Team Name	Codalab Username	F1-score (%)	Accuracy (%)	Precision (%)	Recall (%)
1	TJU-MI	wangxiuxian	63.70	64.89	64.59	63.68
2	TSR (Ray et al., 2025)	ray-sushant	62.91	62.92	64.22	63.07
3		Ryuan	62.80	62.72	64.50	63.10
4	IMU-L	jiaranDiana	61.76	62.13	62.54	61.58
5	MMFusion (Rane, 2025)	prerana3	60.86	61.14	61.23	61.14
6	Multimodal Kathmandu (Maharjan et al., 2025)	Sujal_Maharjan	60.70	61.14	61.18	61.25
7	MLInitiative (Acharya et al., 2025)	ankitbk07	60.59	61.14	60.64	60.59
8	MemeMasters (Shakya and Gurung, 2025)	shrutigurung	60.23	59.96	63.09	60.56
9	Overfitters (Bhattarai et al., 2025)	bidhancb	60.15	60.55	60.27	60.17
10	Silver (Mainali et al., 2025)	rohanmainali	59.30	59.57	59.53	59.47
11	Musafir	MDSagorChowdhury	54.29	54.24	55.55	54.83
12	Luminaries (Esackimuthu, 2025)	akshayyy22	53.05	55.23	54.34	53.55
13	MLP (Verma and Kumar, 2025)	Durgeshverma24iitram	46.74	46.94	49.05	47.23

Table 4: Sub-task C (Classification of Topical Stance) Leaderboard, Ranked by Macro F1-score. All scores are presented as percentages (%). It is to be noted that this leaderboard contains the score till the test deadline and does not consider further runs done by participants as a part of the system description paper.

major confusion between the 'Community' and 'Undirected' classes as a key issue.

Luminaries (Esackimuthu, 2025)fine-tuned ALBERT for multiclass classification and also trained a feedforward ANN. Their ALBERT model achieved an F1-score of 0.4984 and an accuracy of 55.42%, placing 11th on the leaderboard. While the model effectively captured contextual dependencies, the system struggled to distinguish between fine-grained target categories.

wangkongqiang (Kongqiang and Peng, 2025) performed four different benchmarks in different models, where the multinomial naive bayes classification model showed the best, achieving an F1-score of 34.53% and holding 13th position in the leaderboard.

MLP (Verma and Kumar, 2025) employed XLM-RoBERTa + ViT with attention-based fusion. The system achieved an F1-score of 40.96% and an accuracy of 42.17%. Despite leveraging bidirectional cross-modal attention and multiple fusion strategies, the model struggled with fine-grained classification of targets within memes.

4.2.3 Subtask C: Topical Stance Detection

TSR (Ray et al., 2025) applied the FIMIF architecture with low-rank multimodal fusion of compressed text and image embeddings. The model obtained an F1-score of 62.91% and accuracy of 62.92%. The architecture leveraged CLIP embeddings alongside residual layers that emphasized linear relationships, with the multiplicative module capturing more complex feature interactions when beneficial. These results highlight the model's ability to distinguish Support, Oppose, and Neutral stances in multimodal memes with relatively few parameters.

MMFusion (Rane, 2025) adopted an ensemble approach, combining the outputs of three separate multimodal models via probability averaging. The ensemble consisted of a RoBERTa-base with ResNet18, a RoBERTa-base with ResNet34, and a BERT-base with ResNet18, each trained with different random seeds to ensure diversity. This method, which performed better than their initial single-model attempts, used a simple attention mechanism for feature fusion and yielded an F1-score of 60.8%.

Multimodal Kathmandu(Maharjan et al., 2025) introduced a Two-Stage Multiplicative Fusion framework, where CLIP features were projected into higher dimensions, refined through lightweight adapters, and combined via element-wise multiplication. A two-stage fine-tuning procedure stabilized training and improved convergence. Their model achieved an F1-score of 60.70% and accuracy of 61.14%, placing 6th overall.

MLInitiative (Acharya et al., 2025) compared their ResNet-18+BERT and SigLIP2 multimodal models. The SigLIP2 model, which processes image-text pairs in a joint embedding space using a sigmoid-based contrastive loss, obtained better performance. It achieved an F1-score of 60.59%, which resulted in a 7th place ranking on the task leaderboard.

MemeMasters (Shakya and Gurung, 2025) modified their CLIP-based framework by employing a deeper 3-layer classifier head and using a cosine learning rate scheduler. The model found it difficult to distinguish between subtle stance dif-

Rank	Team Name	Codalab Username	F1-score (%)	Accuracy (%)	Precision (%)	Recall (%)
1	TJU-MI	wangxiuxian	78.01	81.07	77.50	78.65
2	TSR (Ray et al., 2025)	ray-sushant	76.83	79.68	76.79	76.87
3	Multimodal Kathmandu (Maharjan et al., 2025)	Sujal_Maharjan	75.29	77.91	75.78	74.91
4	-	paiisfunny	75.16	78.50	74.81	75.57
5	-	Ryuan	74.80	78.30	74.35	75.37
6	MemeMasters (Shakya and Gurung, 2025)	shrutigurung	73.13	75.74	73.86	72.66
7	MLInitiative (Acharya et al., 2025)	ankitbk07	72.88	77.71	71.72	75.26
8	Silver (Mainali et al., 2025)	rohanmainali	72.68	75.94	72.75	72.61
9	-	olivialiudama	71.41	75.35	71.06	71.86
10	-	AkshYat	71.13	73.18	72.75	70.67
11	IMU-L	jiaranDiana	70.31	72.98	71.19	69.85
12	MMFusion (Rane, 2025)	prerana3	65.85	73.37	64.89	70.05
13	MLP (Verma and Kumar, 2025)	Durgeshverma24iitram	65.64	70.02	65.54	65.75
14	Overfitters (Bhattarai et al., 2025)	bidhancb	65.33	72.78	64.46	69.06
15	Musafir	MDSagorChowdhury	62.68	66.07	63.25	62.44
16	Luminaries (Esackimuthu, 2025)	akshayyy22	60.70	68.44	60.30	62.74

Table 5: Sub-task D (Classification of Intended Humor) Leaderboard, Ranked by Macro F1-score. All scores are presented as percentages (%). It is to be noted that this leaderboard contains the score till the test deadline and does not consider further runs done by participants as a part of the system description paper.

ferences, leading to significant misclassifications where both 'Support' and 'Oppose' instances were incorrectly labeled as 'Neutral'. This approach achieved a macro F1-score of 60%.

Overfitters (Bhattarai et al., 2025) implemented their BERTRES model, which leverages a combination of BERT and ResNet-50 embeddings. The system struggled with the complexities of this task, particularly in cases that involved satire or ambiguous sentiment. The model also demonstrated overfitting, which the authors attribute to the imbalanced and sparse label distribution for this specific subtask. Ultimately, the system ranked 9th with an F1-score of 60.15%.

Silver (Mainali et al., 2025) employed their comparative framework of unimodal and multimodal models. Their best system was a CLIP-based model, which achieved a macro F1-score of 59.30%. Even with this top-performing model, the authors reported that it struggled to correctly classify memes containing irony or ambiguous tone. The 'Neutral' class was the most likely to be miscategorized as either supportive or opposing.

Luminaries (Esackimuthu, 2025) trained ALBERT and a feedforward ANN independently for stance classification without ensembling. The ALBERT model achieved an F1-score of 53.05%, ranking 12th overall. The system faced difficulty in interpreting ambiguous or ironic stances, which often rely on subtle linguistic cues.

MLP (Verma and Kumar, 2025) applied their multimodal attention-based fusion models. The best-performing system achieved an F1-score of 46.74% with an accuracy of 46.94%. While the approach captured some multimodal interactions, performance remained limited on subtle stance distinctions.

4.2.4 Subtask D: Intended Humor Detection

TSR (Ray et al., 2025) achieved their strongest results, reporting an F1-score of 76.83% and accuracy of 79.68%. The FIMIF model effectively integrated textual sarcasm with visual cues through its multiplicative fusion layer, enabling the detection of humor and satire in memes. Despite its small size compared to transformer-based multimodal systems, the architecture maintained competitive performance, underscoring the value of low-dimensional fusion for subjective classification tasks.

Multimodal Kathmandu(Maharjan et al., 2025) applied the same Two-Stage Multiplicative Fusion framework, augmented with semantic-aware initialization that seeded classifier weights using CLIP embeddings of descriptive prompts. This system delivered one of the strongest results in the task, achieving an F1-score of 75.29% and accuracy of 77.91%, securing 3rd place overall.

MemeMasters (Shakya and Gurung, 2025) adjusted their CLIP model by incorporating a higher dropout rate and using a class-weighted loss to handle the imbalanced data, which was skewed towards humorous content. The model was conservative in its predictions, often misclassifying humorous content as non-humorous due to the context-dependent nature of online hu-

mor. This system yielded a macro F1-score of 73%.

MLInitiative (Acharya et al., 2025) addressed humor detection using a ResNet-18+BERT fusion model and a more advanced SigLIP2 model. Their results showed that the SigLIP2 architecture was more effective for the task, achieving an F1-score of 72.88%. This performance earned their system the 7th position on the subtask's final leaderboard.

Silver (Mainali et al., 2025) found that multimodal models outperformed unimodal approaches, as visual cues were often critical for contextualizing humor in memes. A CLIP-based model proved to be the most effective, delivering a macro F1-score of 72.68%. Despite this success, the system was prone to making false predictions on content that involved sarcasm or culturally specific jokes that were not conveyed through text.

MMFusion (Rane, 2025) developed a distinct multimodal architecture using DialoGPT-medium for text and ResNet50 for images, choosing DialoGPT for its proficiency with informal, conversational language. Their system applied self-attention to each modality independently before using cross-modal attention and a final gating mechanism to adaptively weight and combine the features. This approach resulted in an F1-score of 65.8%.

MLP (Verma and Kumar, 2025) reported stronger results relative to stance and target identification. Their multimodal architecture reached an F1-score of 65.64% and an accuracy of 70.02%. This indicates that their system was able to capture explicit humorous cues in memes using cross-modal fusion of textual and visual features.

Overfitters (Bhattarai et al., 2025) addressed the humor detection challenge with their BERTRES multimodal fusion model. The model, which integrates text embeddings from BERT and visual features from ResNet-50, found this task to be particularly difficult due to the subjective and culturally specific nature of humor in memes, which made it hard for the model to generalize. Their system achieved an F1-score of 65.33%, resulting in a 14th-place ranking.

Luminaries (Esackimuthu, 2025) utilized ALBERT and an ANN, treating this as a binary classification task. Their fine-tuned ALBERT model achieved an F1-score of 60.70%, ranking 16th overall. Performance was constrained by the subjective and culturally dependent nature of humor, with frequent misclassification of sarcastic or context-heavy instances.

5 Discussion

The results across the four subtasks show the complexities of multimodal analysis of socio-political memes, where humor, satire, and harmful speech often intersect. While multimodal models generally outperformed unimodal baselines, the extent of improvement varied by task, reflecting differences in difficulty, class imbalance, and the interplay of textual and visual cues. Hate Speech Detection (Subtask A) achieved the highest scores, with several teams surpassing an F1-score of 80%, indicating that binary classification of explicit or strongly implied hate is relatively well-handled by current models. In contrast, Target Identification (Subtask B) proved most challenging, with substantial performance drops due to fine-grained labels, skewed class distributions, and frequent overlap between categories such as Community and Undirected.

Stance Detection (Subtask C) showed moderate performance, with top scores in the low 60s, hindered by the difficulty of interpreting sarcasm, irony, and ambiguous sentiment. Humor Detection (Subtask D) fared slightly better, with top teams exceeding 76% F1, suggesting that visual tropes and textual patterns characteristic of humor are more consistently captured by multimodal fusion methods. CLIP-based approaches dominated many leaderboards, while compact, parameter-efficient architectures like TSR's FIMIF (Ray et al., 2025) demonstrated that strong results are achievable with minimal trainable parameters. Attention-based and gating fusion mechanisms yielded mixed benefits, with improvements often dependent on task-specific dynamics.

Persistent challenges include handling subtle, culturally dependent cues, mitigating severe class imbalance, particularly in Subtask B, and resolving multimodal ambiguity when text and image signals conflict or provide weak cues. Future progress will require better handling of fine-grained categories, integration of external knowledge to interpret implicit references, improved balancing strategies, and hybrid architectures that combine precise

language understanding with strong cross-modal alignment. Overall, while current models show promise in detecting overt hate and humor, capturing nuanced communicative intent in marginalized community discourse remains an open challenge.

6 Conclusion

In this paper, we presented the Shared Task on Multimodal Detection of Hate Speech, Humor, and Stance in Marginalized Socio-Political Movement Discourse at CASE 2025 (co-located with RANLP 2025), leveraging the PrideMM dataset, a curated collection of memes related to the LGBTQ+ pride movement. The task was designed to encourage the development of models capable of jointly addressing four interconnected challenges: (i) detecting hate speech, (ii) identifying its targets, (iii) classifying topical stance, and (iv) recognizing intended humor. With participation from 89 registered teams and competitive submissions across all subtasks, the results demonstrated the clear advantage of multimodal approaches over unimodal baselines, while also revealing substantial variation in task difficulty and persistent challenges in handling subtle, implicit, or culturally dependent content. The insights from this shared task provide a valuable benchmark for future research, suggesting the need for methods that combine robust cross-modal integration with cultural and contextual awareness. We hope this work will stimulate continued innovation in multimodal content moderation, contributing to safer and more inclusive online spaces.

Acknowledgments

Funding for part of this research has been provided by the European Union's Horizon Europe research and innovation programme EFRA [grant number 101093026] and ECO-Ready [grant number 101084201].

Broader Impact

This shared task aims to advance multimodal content moderation in contexts involving marginalized socio-political movements, focusing on the LGBTQ+ pride movement. By targeting nuanced, culturally embedded, and often ambiguous content, it encourages the development of fairer, more context-aware AI systems that can mitigate harm while preserving legitimate expression. However, automated moderation must be applied cautiously, as misclassification can silence marginalized voices or misinterpret culturally specific discourse. The PrideMM dataset was curated with rigorous annotation to promote inclusivity and reduce bias, but real-world use should involve human oversight and community input. Beyond moderation, the work offers value for social science, media studies, and policy, supporting safer and more inclusive online spaces while respecting expressive diversity.

References

Ashish Acharya, Ankit BK, Bikram K.C., Sandesh Shrestha, Tina Lama, Surabhi Adhikari, and Rabin Thapa. 2025. MLInitiative at CASE 2025: Multimodal Detection of Hate Speech, Humor,and Stance using Transformers. In *Proceedings of The Eighth Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Abeer AlDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Farhan Amin, Muhammad Abu Horaira, Md. Tanvir Ahammed Shawon, Md. Ayon Mia, and Muhammad Ibrahim Khan. 2025. PhantomTroupe@CASE 2025: Multimodal Hate Speech Detection in Text-Embedded Memes using Instruction-Tuned LLMs. In *Proceedings of The Eighth Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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.

Bidhan Chandra Bhattarai, Dipshan Pokhrel, Ishan Maharjan, and Rabin Thapa. 2025. Overfitters@CASE2025: Multimodal Hate Speech Analysis Using BERT and RESNET. In *Proceedings of The Eighth Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Jamell Dacon, Harry Shomer, Shaylynn Crum-Dacon, and Jiliang Tang. 2022. Detecting harmful online conversational content towards lgbtqia+ individuals. *arXiv preprint arXiv:2207.10032*.

Akshay Esackimuthu. 2025. Luminaries@CASE 2025: Multimodal Hate Speech, Target, Stance and Humor Detection using ALBERT and Classical Models. In *Proceedings of The Eighth Workshop on Challenges and Applications of Automated Extraction of*

Socio-political Events from Text (CASE), Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Akash Gautam, Puneet Mathur, Rakesh Gosangi, Debanjan Mahata, Ramit Sawhney, and Rajiv Ratn Shah. 2020. # metooma: Multi-aspect annotations of tweets related to the metoo movement. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 209–216.

Ali Hürriyetoğlu, Hristo Tanev, Osman Mutlu, Surendrabikram Thapa, Fiona Anting Tan, and Erdem Yörük. 2023. Challenges and applications of automated extraction of socio-political events from text (CASE 2023): Workshop and shared task report. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text*, pages 167–175, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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)*.

Ali Hürriyetoğlu, Surendrabikram Thapa, Gökçe Uludoğan, Somaiyeh Dehghan, and Hristo Tanev. 2024. A concise report of the 7th workshop on challenges and applications of automated extraction of socio-political events from text. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 248–255.

Tomal Paul Joy, Aminul Islam, Saimum Islam, Md. Tanvir Ahammed Shawon, Md. Ayon Mia, and Mohammad Ibrahim Khan. 2025. CUET NOOB@CASE2025: MultimodalHate Speech Detection in Text-Embedded Memes using Late Fusion with Attention Mechanism. In *Proceedings of The Eighth Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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.

Shamika Klassen and Casey Fiesler. 2022. "this isn't your data, friend": Black twitter as a case study on research ethics for public data. *Social Media+ Society*, 8(4):20563051221144317.

Wang Kongqiang and Zhang Peng. 2025. wangkongqiang@CASE 2025: Detection and Classifying Language and Targets of Hate Speech using Auxiliary Text Supervised Learning. In *Proceedings of The Eighth Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Sujal Maharjan, Astha Shrestha, Shuvam Thakur, and Rabin Thapa. 2025. Multimodal Kathmandu@CASE 2025: Task-Specific Adaptation of Multimodal Transformers for Hate, Stance, and Humor Detection. In *Proceedings of The Eighth Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Rohan Mainali, Neha Aryal, Sweta Poudel, Anupraj Acharya, and Rabin Thapa. 2025. Silver@CASE2025: Detection of Hate Speech, Targets, Humor, and Stance in Marginalized Movement. In *Proceedings of The Eighth Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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*.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021a. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796, Online. Association for Computational Linguistics.

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

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*.

Prerana Rane. 2025. MMFusion@CASE 2025: Multimodal Learning for Hate Speech, Target, Stance, and Humor Classification in Marginalized Movement Discourse. In *Proceedings of The Eighth Workshop*

on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE), Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Tabassum Basher Rashfi, Md. Tanvir Ahammed Shawon, Md. Ayon Mia, and Muhammad Ibrahim Khan. 2025. ID4Fusion@CASE 2025: A Multimodal Approach to Hate Speech Detection in Text-Embedded Memes Using ensemble Transformer based approach. In *Proceedings of The Eighth Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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.

Sushant Kr. Ray, Rafiq Ali, and Abdullah Mohammad. 2025. TSR@CASE 2025: Low Dimensional Multimodal Fusion Using Multiplicative Fine Tuning Modules. In *Proceedings of The Eighth Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Kate Scott. 2021. Memes as multimodal metaphors: A relevance theory analysis. *Pragmatics & Cognition*, 28(2):277–298.

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

Shubham Shakya and Shruti Gurung. 2025. Team MemeMasters@CASE 2025: Adapting Vision-Language Models for Understanding Hate Speech in Multimodal Content. In *Proceedings of The Eighth Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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).

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.

Kohtaro Tanaka, Hiroaki Yamane, Yusuke Mori, Yusuke Mukuta, and Tatsuya Harada. 2022. Learning to evaluate humor in memes based on the incongruity theory. In *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, pages 81–93.

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

Durgesh Verma and Abhinav Kumar. 2025. Multimodal Deep Learning for Detection of Hate, Humor, and Stance in Social Discourse on Marginalized Communities. In *Proceedings of The Eighth Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

A Related Work

As memes become a popular way to express opinions on social and political issues, researchers are paying more attention to analyzing both their text and images to detect hate, humor, and ideologically charged content. The Hateful Memes Challenge (Kiela et al., 2020) introduced one of the

earliest benchmark datasets containing synthetic memes with contrastive image-text signals targeting protected categories such as race, religion, and gender. Subsequent datasets like Harm-C and Harm-P (Pramanick et al., 2021a,b) captured real-world political and COVID-related memes, annotated across varying degrees of harmfulness and target categories. CrisisHateMM (Bhandari et al., 2023) similarly focused on memes related to the Russia-Ukraine conflict and included hate speech target classification. Beyond hate detection, other efforts have targeted different aspects of meme communication: Suryawanshi et al. (2020) introduced MultiOFF for offensive meme detection using data from Reddit and Instagram; Tanaka et al. (2022) proposed a humor detection dataset by extracting memes without interpersonal bias; and DisinfoMeme (Qu et al., 2022) focused on disinformation, annotating memes from movements like BLM and Veganism. Parallel efforts have explored multi-aspect datasets capturing a wider spectrum of linguistic phenomena, Gautam et al. (2020) annotated the MeToo movement-related tweets across dialogue acts, sarcasm, stance, and hate; Dacon et al. (2022) labeled LGBTQ-related Reddit comments for toxicity and identity attacks; and Ousidhoum et al. (2019) provided a multilingual corpus annotated for hate, offensiveness, stance, and sentiment. Recent shared tasks have explored these challenges further: the CASE 2024 Climate Activism task (Thapa et al., 2024a) annotated tweets for stance, hate speech, and humor; the NAET dataset (Rauniyar et al., 2023) collected Nepali anti-establishment tweets with multi-aspect labels including satire, hate, and hope speech; and NEHATE (Thapa et al., 2023b) focused on identifying hate speech and its targets in Nepali election discourse. Similarly, the GameTox dataset (Naseem et al., 2025) introduced token-level and intent-level annotations for toxicity in gaming chats. While most prior work focuses on single aspects or monolingual textual analysis, our task offers a multimodal and multi-aspect benchmark covering hate, stance, humor, and targeted hate enabling richer exploration of social discourse through memes.

B Evaluation and Competition

This section outlines the overall framework of our shared task, the evaluation methodology, competition structure, and key logistical information.

B.1 Evaluation Metrics

We employed a suite of standard classification metrics to evaluate performance: accuracy, precision, recall, and the macro F1-score. The official ranking of the participating teams on the final leaderboard was determined based on their macro F1-score.

B.2 Competititon Setup

The shared task was hosted on the CodaLab platform[1], which provided a standardized environment for all participants. The competition was structured into two primary phases: a development phase and a final testing phase, followed by a peer-reviewed paper submission process.

Registration. A total of 89 participants registered for the shared task, which shows strong interest from individuals across diverse professional backgrounds. Geographic diversity was also notable, as indicated by the wide range of email domain affiliations. Of the registered participants, 21 teams submitted their final prediction outputs.

Competition Timelines. The competition officially commenced on April 8, 2025, with the release of the training and evaluation datasets. This initial phase allowed participants a full month to familiarize themselves with the data, develop their models, and perform internal validation.

The final testing phase began on May 8, 2025, with the release of the test set, for which the ground truth labels were withheld. Participants had over two months to apply their systems to the test data, with the testing period concluding on July 12, 2025.

Following the completion of the testing phase, teams were invited to document their methodologies in a system description paper, with submissions due by Aug 4, 2025. These papers underwent a formal review process, and notifications were sent to authors on August 17, 2025. Authors of conditionally accepted papers were given until August 24, 2025, to submit their revised versions, with final notifications sent on August 25, 2025. The camera-ready versions of all accepted papers were due on August 30, 2025. The shared task will culminate with presentations at the CASE Workshop from September 11-13, 2025. This structured timeline, coupled with continuous support for any technical issues, was designed to facilitate a productive and engaging research environment for all participants.

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

Natural Language Processing vs Large Language Models: this is the end of the world as we know it, and I feel fine

Bertrand De Longueville
European Commission
Joint Research Centre
Brussels, Belgium
bertrand.de-longueville@ec.europa.eu

Abstract

As practitioners in the field of Natural Language Processing (NLP), we have had the unique vantage point of witnessing the evolutionary strides leading to the emergence of Large Language Models (LLMs) over the past decades. This perspective allows us to contextualise the current enthusiasm surrounding LLMs, especially following the introduction of "General Purpose" Language Models and the widespread adoption of conversational chatbots built on their frameworks. At the same time, we have observed the remarkable capabilities of zero-shot systems powered by LLMs in extracting structured information from text, outperforming previous iterations of language models.

In this paper, we contend that that the hype around "conversational AI" is both a revolution and an epiphenomenon for NLP, particularly in the domain of information extraction from text. By adopting a measured approach to the recent technological advancements in Artificial Intelligence that are reshaping NLP, and by utilising Automated Socio-Political Event Extraction from text as a case study, this commentary seeks to offer insights into the ongoing trends and future directions in the field.

1 WHAT – the significance of the "ChatGPT revolution" on NLP

To start this commentary in the context of a renowned workshop on Automatic Event Extraction from Text, let's ask ourselves a rather philosophical question: what distinguishes a socio-political event of rather anecdotic importance (e.g. "Donald Trump tweeted he is unhappy about XYZ") from an event that is most likely to mark history (e.g. "WHO Declares COVID-19 a Pandemic and calls to act in consequence")? Natural Language Processing (NLP), our area of research, is unlikely to provide a scientific answer to this question, so I would rather propose an empirical one, a twofold method for measuring the historical significance of a socio-political event. Firstly, it is interesting to observe that individual people witnessing events that make history have very accurate personal memories of what they were doing when it happened. For example, people born in the 1950's or earlier, when being asked what asked they did on 20 July 1969, when Apollo 11 landed on the moon, often provide a precise narrative about their activities, even decades after the fact. There is a second method to recognise key events: they render obsolete almost immediately common beliefs and thoughts that held authority before them. For example, the permanence of USSR as a political entity could have been considered as obvious for most of its citizens … until the collapse of the Berlin Wall, and subsequently of the PCUS regime a couple of years later.

I am always reluctant to use the expression "ChatGPT revolution" to designate the hype that followed, in the fall/winter 2022-2023, the launch of OpenAI's conversational AI Chatbot powered by the GPT3.5 Large Language Model. After all, Generative Pretrained Transformer (GPT) Models are the fruit of a decades-long continuum of technical evolution, from Support Vector Machines to Deep Learning, from early efforts to perform statistical machine translation to massive training of general-purpose language models (Johri et al., 2021). If we look at the performance of NLP applications we, experts, must acknowledge that the turning point has probably occurred several years earlier, with the discovery of the Transformer architecture, unlocking the efficiency of machine learning models for natural language understanding, thanks to the mechanism of attention (Vaswani et al., 2017). But for the general public, the

Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025, pages 32–37, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-004

"revolution" has happened when such transformers became conversational. I must admit that when I saw GPT 2.0 generating fake research paper reviews (or even fake papers) (Bartoli & Medvet, 2020), I was not overly impressed. In my view, natural language understanding was where societally relevant use cases resided, not in the generation of ersatz human texts.

I dismissed generative LLMs, seeing them as useless stochastic parrots (Bender et al., 2021)… and the entire world proved me wrong. Of course, it is the nature of a hype is to feed itself. And the impressive uptake of ChatGPT[1] can be explained by cultural factors, rather than by its technical innovation: the myth of the talking machine, from Medieval tales of the Brazen Head to "2001: a space Odyssey" … But this is not my point. Let's simply take note that the advent of LLMs matches our twofold criteria and therefore qualifies as a "significant" historical event. Firstly because, if you ask colleagues and friends, most will be able to tell how they encountered for the first time an LLM-powered Chatbot (often, without knowing it was LLM-powered). Personally, I recall precisely the circumstances in which my hierarchical superior explained me (gently, but firmly) that, as the Head of a Text Mining Competence Centre, I could not ignore the advent of conversational AI, "as a matter of existential threat to my research team". And secondly because it made obsolete many widespread claims about AI and Language Technologies. Take for example the widely cited and seminal paper in our area of interest, from 2016 and titled "Growing pains for global monitoring of societal events" (Wang et al., 2016) : it claims – rightfully, then – "the text-processing systems used in event coding are still similar to ones developed more than 20 years ago". Could we say this about our event extraction systems in 2025? I do not think so.

2 SO WHAT – LLMs as "game changers" for NLP

In this context, we may wonder: are LLMs truly game changers for Natural Language Processing in general, and for Automated Event Extraction from text in particular? An abundant literature suggests so, which corroborates the intuitions shared in the previous section (Cronin, 2024; Törnberg, 2023; Yang et al., 2024). Let's reflect further, from an NLP practitioner's perspective, on the implications of General-Purpose LLMs for Automated Event Extraction.

At first glance, one may claim we are reaching "the end of history" (Chernyavskiy et al., 2021) for NLP… after all, LLMs act as remarkably versatile zero-shot machine learning models, being able to extract almost any relevant information from a piece of text, relying on almost human-level of text understanding in hundreds of natural languages, and on a "world model" derived from their training on a significant share of all human knowledge ever produced (in the form of millions of books, encyclopaedias, scientific articles, websites, conversations, blogs, etc.). So, "game over" for NLP scientists, let's all retrain as "prompt engineers" by practicing the art of asking the right question to General Purpose LLMs/oracles…

Well, it's not that simple.

First of all, let's not forget the inference cost aspects. In Socio-Political Event Extraction, real-life use cases often require the processing of vast amounts of raw text (typically, news articles or field reports), so the computing power to process them in near-real time can become a significant bottleneck. Based on my own experience, I would say there is a ratio of about 1 to 50, or even 100, in terms of computing power required to run a "good old" BERT-like model compared to state-of-the-art LLAMA 4 or Mistral 3.1 open weights models. Moreover, the latest models require costly and powerful GPU hardware cards that are on high demand, while BERT-like models run on older hardware that is likely to be already amortised in terms of cost, and more easily available for purchase. Literature shows that properly fine-tuned models of the BERT generation perform at very high levels for specialised tasks such as geocoding(Tanev & De Longueville, 2023), sentiment analysis (Di Nuovo et al., 2024), discourse analysis (Stefanovitch, De Longueville, et al., 2023), or topic mining (Stefanovitch, Jacquet, et al., 2023), which are all relevant for Automated Event Analysis purposes. So one may wonder: why would we need to invest in a Ferrari when we have a highly adaptable fleet of Land Cruisers at hand?

There is another reason why LLMs are not "the end of history" for NLP. If LLMs can provide an

[1] ChatGPT reached 100 million users in two month, while it took Instagram two years to reach this symbolic step (Deng et al., 2023)

answer to virtually any question, it is never guaranteed that such an answer – although remarkably crafted from a linguistic point of view – is factually correct. The problem of hallucinations is well known and widely discussed (Huang et al., 2025), but interestingly, the root causes of such behaviour are often overlooked. One of these reasons is sycophancy[2] (Malmqvist, 2024). The need to provide an answer at any cost, in order to please the interlocutor is deeply embedded in the LLM's training process, as their reward function includes some form of "user satisfaction". For this reason, even the best prompt in the world cannot completely avoid sycophantic behaviour and hallucinations. So when facts matter, like in NLP and a fortiori in its Automated Event Extraction use cases, LLMs can never be blindly trusted.

Another trustworthiness issue with LLMs is linked to their "knowledge" component: because they are so eloquent, and because they have been trained on much more information than we could possibly read in our entire lives, we assume they are almost omniscient. But in fact, the world knowledge they seem to feature is more a by-product of their next-word-prediction ability than the result of an accurate and fit-for-purpose world model. LLMs talk, they know and they even reason… but not in the exact same way we do (De Longueville et al., 2025). It is easy to arrive at a misunderstanding situation with LLMs; in brainstorming or creative use cases, that can even be an advantage. But in NLP, where the goal is precisely to extract accurate information from inherently ambiguous natural language, misuse of LLMs abilities can lead to disappointment.

To overcome the "knowledge" ambiguity of LLM's behaviours, the best solution resides in the engineer ever more complex systems that feed them with the right contextual knowledge, in a process called Retrieval Augmented Generation (Lewis et al., 2020). In the context of event extraction, a RAG pipeline can for example include some Gazetteer lookup to improve geocoding (Tanev & De Longueville, 2023).

But if LLMs "know", they also "reason": imagine a sentence like "the political meeting will take place in Zoom". A RAG-enabled AI system, designed to rigorously lookup places in a comprehensive gazetteer would probably geocode such an event in Zoom, a Village in Soreng Tehsil in West District of Sikkim State, India (lat 27.144465910353176, long 88.26329879818186), while we, humans, would rather infer "Zoom" designates an online video-conferencing platform[3]. This example shows that the "reasoning" component of LLMs cannot be blindly trusted either, even when LLMs are fed with the best data and follow the best-crafted prompt instructions. It is important to have that concept in mind, as NLP experts, since with the advent of Agentic AI systems (Chawla et al., 2024), we will increasingly rely on LLM's ability to reason.

Based on the above, one may conclude that since LLMs are not magically addressing any possible issue, then they are junk… Since we cannot trust 100% for a task, then we cannot trust them for any task – including our preferred one: extraction of spatiotemporal information patters from text. This would not be a rational approach to the promises of LLMs. As scientists, we should wonder: if I cannot trust 100% my LLM system for this task, then to what percent can I trust it? And there, we start thinking in terms of precision and recall … There we go again!

3 NOW WHAT – the new NLP that looks like the good old one

Everything changes, but nothing changes: on the one hand LLMs can act as prodigious zero-shot information extraction machines that open new perspectives for NLP applications, but on the other hand, their precision and recall need to be accurately measured…

Evaluating the performance of NLP software modules to perform specific tasks, like automatically extracting information about socio-political events from text is a classic activity for NLP scientists. It requires the creation and curation of "gold standard" corpora, where the expected outcome of a large number of instances of the same task is encoded (usually, by human annotators), and on which variants of the software module are tested, until the highest possible F-score (Derczynski, 2016) is reached, expressing the best possible compromise between precision and recall.

So, really, nothing new under the sun for NLP practitioners.

[2] According to the Cambridge Dictionary, sycophancy is defined as the "behavior in which someone praises powerful or rich people in a way that is not sincere, usually in order to get some advantage from them".

[3] This is not a fictitious case: I saw it happening.

However, understanding the underlying reasons for LLMs successes and failures to provide results corresponding to gold standards opens new research perspectives. For example, there is a need to better understand and assess spatiotemporal reasoning abilities of AI systems based on LLMs, and how formal ontologies (like gazetteers for place names, or named entities databases) can complement LLM's internal (and perfectible) world model for tackling hallucinations and supporting entities disambiguation. In other words, there is a need to further explore hybrid approaches aiming at developing NLP processing pipelines that involve LLMs, advanced Retrieval-Augmented Generation technique and more deterministic approaches like rule-based on symbolic AI components. Interesting developments have been recently published that go in that direction for geoparsing (Halterman, 2023) or epidemic events detection (Consoli et al., 2024). These are examples to follow while exploring other epistemic tasks related to event extraction..

Also, while the founding principles of NLP task-specific evaluations remain valid, the scientific methods to measure the efficiency of non-deterministic AI pipelines executing complex event-extraction processes remain to be studied, thus paving the way for next-generation socio-political event extraction research. Inspiration could come from similar – but distinct – language technology research areas. For example, studying how to improve the knowledge extraction component of an LLM pipeline with a carefully engineered RAG component (Ceresa et al., 2025), or by developing integrated "software + datasets" bundles to well targeted task evaluation of specialised NLP software packages (Bassani & Sanchez, 2024).

To achieve scientifically reproducible results in this novel area of research for LLM-ready NLP, it is essential that academic organisations have the ability to run their own LLM inference systems. With the trend of ever larger LLMs[4], and given the IT infrastructures constraints described above, there is an increasing trend to rely on LLM-as-service provided through Application Programming Interfaces. This creates an additional difficulty for scientists, as such models, often provided commercially, do not fully disclose their detailed systems specifications (e.g. input filters, output filters or system prompts which have a proven strong impact on an AI service behaviour (De Longueville et al., 2025), and may change without prior notice, making previous NLP task evaluations obsolete. As a consequence, LLM-as-service is even more a black box than any Deep Learning model, as the LLM itself is surrounded with undisclosed technical components that influence its output.

It is thus a matter of independent science – and ultimately of Sovereignty – that Academic organisations remain capable of fully controlling the execution environment of the LLMs they base their research on. The availability of state-of-the-art open weight LLMs is therefore crucial for academia, and will become of paramount importance as the advent of Agentic AI will introduce novel paradigms for knowledge workers, and among them scientists especially, for interacting with data and information, using AI systems as "mediators" (e.g. when using an LLM-powered tool to perform systematic literature reviews).

In the light of the above, we may draw this oxymoronic conclusion: for NLP, the advent of general purpose LLMs is both a revolution and an epiphenomenon.

Been there, seen that: as a geospatial scientist, I saw in the early 2000's the combination of cheaper GPS devices, pervasive Internet connections and web 2.0 technologies like AJAX lead to a paradigm shift in my research area (De Longueville et al., 2010). The release of the Google Earth to the wide public in 2005 embodied this revolution for the general public and created a hype similar to the one around ChatGPT nowadays. Faced with such technologies enabling interoperable analysis and visualisation of geospatial data on a smooth Digital Earth interface, some may have wondered: is it the end of history for geospatial sciences? Yet, this research area remains vibrant 20 years later, increasing our Earth Observations capabilities and refining our common understanding of complex planetary phenomena.

Will the same happen to NLP with the advent of LLMs and Agentic Systems in the 2020's? In other words, dear NLP scientists, are you ready to cope with the rollout at large scale of "GPS and Digital Earth, but for the knowledge"? Your answers to these questions will shape the future of NLP in the next decades.

[4] Although this trend is perceived as plateauing already (Villalobos et al., 2024), the size of current top-performing models (which is only based on assumptions for commercial, non-open-source models) already exceeds the IT infrastructure capacity of most Universities and Research Centres for running them at large scale for inference.

References

Bartoli, A., & Medvet, E. (2020). Exploring the Potential of GPT-2 for Generating Fake Reviews of Research Papers. In *Frontiers in Artificial Intelligence and Applications*. IOS Press. https://doi.org/10.3233/faia200717

Bassani, E., & Sanchez, I. (2024). Guardbench: A large-scale benchmark for guardrail models. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 18393–18409.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Ceresa, M., Bertolini, L., Comte, V., Spadaro, N., Raffael, B., Toussaint, B., Consoli, S., Piñeiro, A. M., Patak, A., Querci, M., & others. (2025). Retrieval Augmented Generation Evaluation for Health Documents. *arXiv Preprint arXiv:2505.04680*.

Chawla, C., Chatterjee, S., Gadadinni, S. S., Verma, P., & Banerjee, S. (2024). Agentic AI: The building blocks of sophisticated AI business applications. *Journal of AI, Robotics & Workplace Automation*, *3*(3), 1–15.

Chernyavskiy, A., Ilvovsky, D., & Nakov, P. (2021). Transformers: "The End of History" for Natural Language Processing? In *Lecture Notes in Computer Science* (pp. 677–693). Springer International Publishing. https://doi.org/10.1007/978-3-030-86523-8_41

Consoli, S., Markov, P., Stilianakis, N. I., Bertolini, L., Gallardo, A. P., & Ceresa, M. (2024). Epidemic Information Extraction for Event-Based Surveillance Using Large Language Models. *International Congress on Information and Communication Technology*, 241–252.

Cronin, I. (2024). *Decoding large language models: An exhaustive guide to understanding, implementing, and optimizing LLMs for NLP applications*. Packt Publishing.

De Longueville, B., Annoni, A., Schade, S., Ostlaender, N., & Whitmore, C. (2010). Digital Earth's Nervous System for crisis events: Real-time Sensor Web Enablement of Volunteered Geographic Information. *International Journal of Digital Earth*, *3*(3), 242–259. https://doi.org/10.1080/17538947.2010.484869

De Longueville, B., Sanchez, I., Kazakova, S., Luoni, S., Zaro, F., Daskalaki, K., & Inchingolo, M. (2025). *The Proof is in the Eating: Lessons Learnt from One Year of Generative Ai Adoption in a Science-for-Policy Organisation*. https://doi.org/10.2139/ssrn.5141665

Deng, Y., Zhao, N., & Huang, X. (2023). Early ChatGPT User Portrait through the Lens of Data. *2023 IEEE International Conference on Big Data (BigData)*, 4770–4775. https://doi.org/10.1109/bigdata59044.2023.10386415

Derczynski, L. (2016). Complementarity, F-score, and NLP Evaluation. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 261–266). European Language Resources Association (ELRA). https://aclanthology.org/L16-1040/

Di Nuovo, E., Cartier, E., & De Longueville, B. (2024). Meet XLM-RLnews-8: Not Just Another Sentiment Analysis Model. In *Natural Language Processing and Information Systems* (pp. 24–35). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-70242-6_3

Halterman, A. (2023). *Mordecai 3: A Neural Geoparser and Event Geocoder*. https://arxiv.org/abs/2303.13675

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, *43*(2), 1–55. https://doi.org/10.1145/3703155

Johri, P., Khatri, S. K., Al-Taani, A. T., Sabharwal, M., Suvanov, S., & Kumar, A. (2021). Natural Language Processing: History, Evolution, Application, and Future Work. In *Lecture Notes in Networks and Systems* (pp. 365–375). Springer Singapore. https://doi.org/10.1007/978-981-15-9712-1_31

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-

Augmented Generation for Knowledge-Intensive NLP Tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 9459–9474). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

Malmqvist, L. (2024). *Sycophancy in Large Language Models: Causes and Mitigations*. https://doi.org/10.48550/ARXIV.2411.15287

Stefanovitch, N., De Longueville, B., & Scharfbillig, M. (2023). TeamEC at SemEval-2023 Task 4: Transformers vs. Low-Resource Dictionaries, Expert Dictionary vs. Learned Dictionary. *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 2107–2111. https://doi.org/10.18653/v1/2023.semeval-1.290

Stefanovitch, N., Jacquet, G., & De Longueville, B. (2023). Graph and Embedding based Approach for Text Clustering: Topic Detection in a Large Multilingual Public Consultation. *Companion Proceedings of the ACM Web Conference 2023*, 694–700. https://doi.org/10.1145/3543873.3587627

Tanev, H., & De Longueville, B. (2023). Where "where" Matters: Event Location Disambiguation with a BERT Language Model. In A. Hürriyetoğlu, H. Tanev, V. Zavarella, R. Yeniterzi, E. Yörük, & M. Slavcheva (Eds.), *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text* (pp. 11–17). INCOMA Ltd., Shoumen, Bulgaria. https://aclanthology.org/2023.case-1.2/

Törnberg, P. (2023). *How to use LLMs for Text Analysis* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2307.13106

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. https://doi.org/10.48550/ARXIV.1706.03762

Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., & Hobbhahn, M. (2024). Position: Will we run out of data? Limits of LLM scaling based on human-generated data. *Forty-First International Conference on Machine Learning*.

Wang, W., Kennedy, R., Lazer, D., & Ramakrishnan, N. (2016). Growing pains for global monitoring of societal events. *Science*, *353*(6307), 1502–1503. https://doi.org/10.1126/science.aaf6758

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Zhong, S., Yin, B., & Hu, X. (2024). Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. *ACM Transactions on Knowledge Discovery from Data*, *18*(6), 1–32. https://doi.org/10.1145/3649506

Machine Translation in the AI Era: Comparing previous methods of machine translation with large language models

William Jock Boyd
School of Computing
and Communications
Lancaster University
United Kingdom
williamboyd106@gmail.com

Ruslan Mitkov
School of Computing
and Communications
Lancaster University
United Kingdom
r.mitkov@lancaster.ac.uk

Abstract

The aim of this paper is to compare the efficacy of multiple different methods of machine translation in the French-English language pair. There is a particular focus on Large Language Models given they are an emerging technology that could have a profound effect on the field of machine translation. This study used the European Parliament's parallel French-English corpus, testing each method on the same section of data, with multiple different Neural Translation, Large Language Model and Rule-Based solutions being used. The translations were then evaluated using BLEU and METEOR scores to gain an accurate understanding of both precision and semantic accuracy of translation. Statistical analysis was then performed to ensure the results validity and statistical significance. This study found that Neural Translation was the best translation technology overall, with Large Language Models coming second and Rule-Based translation coming last by a significant margin. It was also discovered that within Large Language Model implementations that specifically trained translation capabilities outperformed emergent translation capabilities.

1 Introduction

This study aims to compare previous and current methods of Machine Translation (MT) with Large Language Models (LLMs) to gauge the effectiveness of novel technologies in the field of MT.

The continuous improvement of technology in the MT space often leads to older methods being left behind – especially in the modern day as more and more companies make the pivot to LLMs. These previous methods, such as rules-based MT, can be effective in situations where there is a lack of resources available to train models. Additionally, LLMs trade off of accuracy for natural sounding translations could cause myriad issues in areas where accuracy is paramount such as the medical or legal fields. This suggests that Neural Machine Translation (NMT) is a better solution for these contexts at the moment. The question of LLMs in the field on MT is still in the early stages of being researched but does have promising results in early studies (Mujadia et al., 2023). However, these studies are often made with comparisons to other LLMs, giving no context as to how they perform against other methods of translation. Given the intense amount of resources required to train and run LLMs, an accurate inter-method comparison would allow potential users of these systems to evaluate the pros and cons before committing the time and resources required to use them.

This paper aims to create a clear picture on how rules-based translation, neural translation, and large language models and compare to each other on translation of the same text, and how different implementations of these methods can affect translation quality. This will give future research a baseline to compare from when progressing the field. This research is novel in that no other study has ever compared these three methods of MT in the same framework before. These new contributions will provide a clear picture of the current MT landscape giving insight as to where research should go in future. They will also let developers planning to use MT as part of their product to make an informed decision on which method is best for them, based on the trade-offs of each one. Within this study the efficacy of different translation approaches for LLMs will also be investigated, allowing developers of this technology to tailor their efforts depending on the task.

This study is highly relevant to the automatic extraction of socio-political events from text, given its focus on automatic translation and multilinguality. In multilingual contexts, translation methods are often essential for enabling such extraction. The data used in this research are drawn from the European Parliament's French–English parallel corpus, which provides extensive coverage of socio-

Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025, pages 38–51, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-005

political events. The findings of this research offer valuable guidance to researchers in selecting suitable approaches for tackling multilingual tasks of this kind.

2 Related Work

Although significant research on each of these methods has been done individually. And research on comparisons between systems like RBMT, NMT and LLMs has been done, these systems have never all been compared together using the same corpus with the same preprocessing on the translation results. Additionally, a majority of research only compares two types of systems at a time, whereas this study compares 3 types of systems with different implementations of those types. The rest of this subsection will discuss prior studies done on this topic, the limitations of that research and the significance of the research being done in this study.

Historical studies naturally covered RBMT and Statistical Machine Translation (SMT) systems such as this study by Costa-Jussà et al. (2012) comparing RBMT and SMT on Catalan – Spanish MT systems across 2 domains. This research was key in defining performance differences between systems. Another key paper evaluating direct performance comparisons between the two systems is the paper by S and Bhattacharyya (2017) which uses the Marathi–Hindi language pair. This is a study with very different takeaways due to the structural differences between Marathi and Hindi, compared to the very similar languages of Catalan and Spanish. In more modern research NMT models have started to be included as part of these studies with multiple studies being published in comparing all 3 system types by De Silva and Hansadi (2024) and Dwivedi et al. (2025) covering this area of research. Additionally, as LLMs have started displaying more and more translation capabilities comparison with existing NMT solutions has started to be done. Such as a paper by Sizov et al. (2024) comparing NMT, LLM, and human translations using human and automatic evaluation. This study sets itself apart by comparing the translations LLMs produce to other systems outputs, rather than focusing on the technique specifically used to get the LLMs to produce this output. However, all of these studies have limitations which will be addressed in the next section.

In studies done less recently only two different methods were compared, not allowing a complete and fair comparison across multiple different systems. This aspect did change with the advent of NMT as researchers wanted to see how it would match up with pre-existing techniques. After the introduction of LLMs to the MT space this focus has narrowed again as studies look to see how LLMs match up against the latest and greatest technologies on offer, rather than how they fit amongst all the available technologies. Additionally according to a meta-analysis by Marie et al. (2021) BLEU scores have been used ineffectively. As studies often copied scores directly from other research without any consideration for how the score was calculated, rendering the comparison invalid. Additionally, without statistical significance testing, the difference between the two scores could be completely coincidental, this is an important tool that is rarely used and the usage of which has been declining over time.

3 Methodology

This section will cover the design of the experiment carried out, including the questions to be answered by the experiment; the translation systems being evaluated and any configurations required to make them work; the corpus these translation systems were tested on; the evaluation metrics use; their specific implementations and the statistical analysis methods used. The experiment protocol will then be discussed, with an evaluation of how these protocols ensure fair comparison and an explanation of how the scores were calculated and compared.

3.1 Corpus Selection and Preparation

The corpus used was the European Parliament's French-English parallel corpus (Koehn, 2005). This was chosen as it covers a variety of domains with discussions ranging from law to the medical field, to nature conservation. This variety enables an excellent insight into how MT systems perform across multiple domains. In addition, the size of this corpus allows for ample development and experiment sets, meaning the development of the testing systems can emulate the experiment itself more closely in terms of scale, without restricting the size of the experiment data. The first 40,000 lines of the last 10% of the data were used as a development set to ensure the integrity of the data, then the next 60,000 lines made up the experiment data set. The only preprocessing done on the data set was to remove unreasonably long sentences that would

exceed the token limits of the models being used.

3.2 Rule-based Model

Given the lack of freely available rule-based models, the only model evaluated in this study is Apertium (Forcada et al., 2011), an open-source RBMT toolkit. For this study version 2.9.4 of the base Apertium model, the English Apertium version, and the French Apertium version were installed. Then the French English language data from the Github was downloaded and the instructions there were used to install and set up the pair. To access the system the command line was used running a shell script that would split the complete experiment file into chunks. Apertium would process each chunk then the translations would be recombined in order. Apertium was chosen in this study as it is the most accessible RBMT model and has been used in multiple research studies previously (Costa-Jussà et al., 2012); (Corbí-Bellot et al.).

3.3 Neural models

Three neural models were assessed in this study to allow different styles and implementations to be evaluated against LLMs, enabling a better overall picture of how they fit in the space. All models were run locally with Huggingface's transformers library (Wolf et al., 2020) using the pipeline interface in Python to send data to the models and receive outputs. The largest models possible were used, as generally the larger the model, the better it performs. Every model was used in the default configuration, with the source languages being specified as French and the target language as English. The neural models chosen as part of this study are the following:

The Marian NMT system is a purely NMT system that uses the transformer architecture (Junczys-Dowmunt et al., 2018), it was developed as an efficient C++-only implementation of the architecture detailed in the paper "Attention is All You Need" (Vaswani et al., 2017). The particular version used was the French-English model from Opus MT (Tiedemann and Thottingal, 2020); (Tiedemann et al., 2022), which is a Marian model trained on the Opus parallel corpus.

Meta's M2M100 model (Fan et al., 2020) is a multilingual translation model that supports translation across 100 different languages. It still uses the same attention mechanism proposed by Vaswani et al. but only requires one model to translate between all these languages. M2M100 was created to address the traditional "English-Centric" approach of multilingual translators, which typically involves translating the source language into English, then English into the target language. The version used in this study was the 1.2 billion parameter version in order to enhance accuracy.

Meta's No Language Left Behind (NLLB) model (Team et al.) is another multilingual translation model but it supports many more languages. NLLB supports 200 different languages, with 150 of them being low-resource languages. The specific model used in this study is the 1.3 billion parameter version, the goal was to use the 3.3 billion parameter version but due to computing resource constraints, this option could not be used.

3.4 Large Language Models

Two LLMs were evaluated in this study to assess how different approaches towards translation capabilities in LLMs can change their effectiveness. Both models were run locally using Huggingface's transformers library and pipeline interface. The LLMs chosen for evaluation are the following:

Google's Text-to-Text Transfer Transformer or T5 (Raffel et al., 2023) is a large language model that treats every NLP task as a text-to-text problem[1]. This approach means T5 can in effect switch modes; this allows the system to approach translation as a task it was directly trained for, rather than as an emergent capability. The uniqueness of T5's approach positions it in a middle ground between NMT systems that can only translate and LLMs that are not trained for translation whatsoever. This technique significantly improves T5's ability to follow instructions and perform zero-shot tasks, allowing T5 to perform in this study despite the constrained computing resources. The specific model version was the FLAN-T5 large, an instruction-tuned version of T5. The model was used in its default configuration with the maximum number of new tokens it was allowed to produce set to 256. When translating, the model was prompted with "Translate from French to English" followed by the sentence to be translated.

Meta's Large Language Model Meta AI (Llama) (Touvron et al., 2023) is an open-source family of LLMs that aims to democratise AI access and enable research advancement. They are a more

[1]Many researchers consider T5 a Deep Learning model not an LLM. For the purposes of this study, T5 will be classed as an LLM due to its generation capabilities. Additionally, the use of T5 large gives more LLM like behaviour.

standard style of LLM being decoder only and pre-trained on large text corpora, meaning translation is an emergent capability. The Llama version used in this study was Llama 3.2 instruct with 3 billion parameters (Grattafiori et al., 2024). The instruct version is fine-tuned on instruction following data, this will improve the model's adherence to the translation request but not the translation itself. The configuration of the model was set to a maximum of 300 new tokens the precision of the model had to be reduced from 32-bit to 16-bit due to resource constraints. This causes a small reduction in overall accuracy, particularly in more nuanced expressions, but is necessary given the constraints of the experiment. The model was also set to only return the response to the prompt. To prompt the model, lists of dictionaries with role and content sections were used. The prompt used was "You are a French to English Translator, translate the input sentences and only give the output sentence" in the system role to set up the model, then in the user role the sentence was given to the model to be translated.

3.5 Evaluation Metrics

Two automated evaluation metrics were used in this study, BLEU score (Papineni et al., 2002) and METEOR score (Banerjee and Lavie, 2005). This approach was used as BLEU score alone can lead to incorrect conclusions about which systems are better according to a meta-evaluation of MT research (Marie et al., 2021), using METEOR avoids this pitfall and also evaluates the systems from a semantic perspective. The Python Natural Language ToolKit (NLTK) (Bird et al., 2009) implementations of both these scores were used. The reference translations for systems to be evaluated against were taken from the Europarl parallel corpus and no modifications were made to the reference translations.

To calculate BLEU score for each translation, the score for each sentence was calculated using the sentence_bleu() function in NLTK, then all the scores were averaged to get an overall score for the translation. Each n-gram was weighted equally, and no smoothing function was used. Sentence level BLEU calculation was used so that bootstrapping could be performed as part of the statistical analysis.

To calculate the METEOR score for each translation, the score for each sentence was calculated using the single_meteor_score() function as there was only one hypothesis per reference translation. The default parameter settings for this implementation were used as they have been studied and calibrated to align with human judgements.

3.6 Data

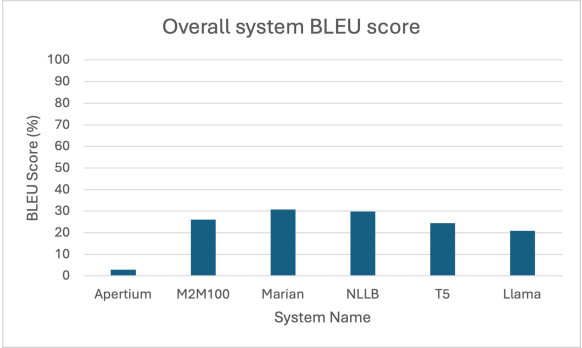

Figure 1: Graph of overall BLEU score for each system

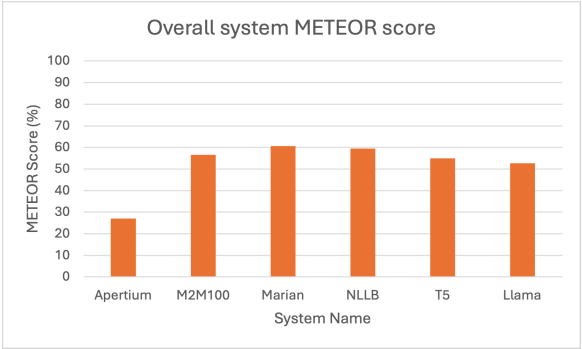

Figure 2: Graph of overall METEOR score for each system

System name	BLEU Score	METEOR Score
Apertium	2.86	27.07
M2M100	26.03	56.49
Marian	30.8	60.59
NLLB	29.8	59.44
T5	24.47	54.97
Llama	20.88	52.6

Figure 3: Table of overall BLEU and METEOR scores for each system

3.7 Statistical Analysis

In order for any conclusions to be made about the results their statistical significance needs to be evaluated to demonstrate they are not just random chance. A meta-evaluation Marie et al. (2021) found that only a minority of papers perform statistical significance testing on their scores. This study addresses this methodological gap by using

bootstrap resampling to ensure the significance of the scores.

Bootstrap resampling was used to create 1000 samples for each system by randomly selecting the sentence level scores from each system with replacement. The size of each sample was 60000 scores - the same size as the original dataset for each system. The overall average of each sample was then recorded so that distributions of these averages could be made and evaluated for each system.

From these distributions, the mean, standard deviation, 95% confidence intervals, minimum, and maximum were calculated for each system. The mean is the primary performance metric and the number that will be compared between systems. The standard deviation shows the variability across samples and how consistent each system's performance is. The 95% confidence intervals establish a range where the true score for each system likely exists. The minimum and maximum values show the best and worst bootstrap samples - a small gap between these two numbers indicates less variability. Together, these metrics give a comprehensive statistical profile of each system's performance. Excel's standard functions were used to calculate these metrics;

The AVERAGE() function was used to calculate the mean of the bootstraps for each system.

STDEV.S() was used to calculate the standard deviation.

The 95% confidence intervals were calculated using the CONFIDENCE.NORM() function, which determines the margin of error based on standard deviation, alpha level, and sample size. An alpha level of 0.05 indicates a 95% confidence interval.

The MIN() and MAX() functions were used to calculate the minimum and maximum sample values for each system.

To compare each system with every other system meaningfully, the p-values between each system were calculated. To calculate the p-values the official result between two systems is compared with every bootstrap sample to see if they match, i.e. if in the official results one system is better; is it better in each bootstrap sample? If less than 5% of the bootstrap results contradict the original finding, meaning $p < 0.05$, then the official result is statistically significant. Calculation of these values was done in excel. To calculate the p-values between systems a formula was implemented to count the number of occurrences where the bootstrap result matched the actual comparison result between the two systems. A sum of these occurrences was then done, that sum was divided by 1000 and taken away from 1 to get the final value. The formula is as follows:

1 - SUM(IF(system 1 bootstrap values > system 2 bootstrap values, 1, 0))/1000

3.8 Apertium Translation

As Apertium cannot handle a huge number of lines at once, the translation data was split into chunks of 150 lines with each chunk in its own file. Apertium was then given each line from each file to translate, every 50 sentences the translations would be written to a file, giving 3 new translated files for every chunk. This was done for every file, and then the files were recombined to create a sentence-aligned file containing every translated sentence.

3.9 Neural Translation

The neural translation was all done from within one file, with each translator translating the source file sequentially so each could have the maximum compute resources available to it. Each system was set up using Huggingface's pipeline in translation mode, and then a for loop iterating through each line in the file was started, yielding each line to the pipeline, the corresponding output was then written to the results file in the same order as the source file ensuring sentence alignment.

3.10 LLM Translation

The LLM translation was done from two separate files given their need for slightly different setups and prompt structuring. T5 was also implemented with Huggingface's pipeline module, it was set up in text-to-text generate mode, with the number of max new tokens allowed to be generated each time set to 256 due to memory constraints. The same for loop to iterate through each line in the source file was used, the input to the model was "Translate from French to English *sentence to translate*". For Llama's implementation, an identical process was used, however, Llama was set up with a max new token count of 300 and 16-bit precision.

3.11 Score Calculation

To calculate the overall METEOR and BLEU scores for each system the individual score of both types for each sentence was calculated then an average of all these sentences was calculated to get

the overall score for each system. To perform bootstrapping a score was randomly selected from the original population of the 60,000 sentence-level scores and added to a new sample but left in the original population. The overall average for the sample was then calculated and added to a list of bootstrap averages for that system.

3.12 Summary

The comprehensive, robust approach detailed in this chapter shows that this experiment is competently able to answer the research questions posed. With a strong framework designed to effectively evaluate each method against the other, using multiple implementations of methods to gain a comprehensive understanding of the performance of each. The use of the Europarl corpus provides a diverse and well-established dataset for translation tasks. The use of both BLEU and METEOR provides a more thorough analysis of the translation quality of each system, as one evaluates the accuracy of the sentence and the other evaluates the semantic meaning. Additionally, statistical analysis with bootstrapping validates the significance of these results - ensuring that conclusions drawn from this study are reliable.

4 Discussion

This section will analyse the results presented in the previous section and discuss their implications.

4.1 Primary Research Question 1

How do RBMT, NMT, and LLM translation approaches compare across general translation tasks in the French-English language pair?

The initial comparison is quite clear with NMT coming out on top with the highest performing NMT system, Marian, having a BLEU score of 30.8% and a METEOR score of 60.59% (3), NMT was then followed by LLMs with the highest scoring LLM, T5, having a BLEU score of 24.47% and a METEOR score of 54.97% (3). RBMT was then last with a significantly larger gap as Apertium had a BLEU score of 2.86% and a METEOR score of 27.07% (3). This huge gap of nearly 28% in BLEU and nearly 33% shows the significant advancements that have been made in the space since the creation of Apertium. In particular, the larger disparity in METEOR scores shows NMT's ability to maintain semantic coherence over the whole translation compared to RBMT. Given NMT's dominance in the study, a comparison between them provides insight into which implementation provides the best translation. The best system was Marian, followed by NLLB with a BLEU score of 29.8% and a METEOR score of 59.44% (3) with M2M100 coming last in the category with a BLEU score of 26.03% and a METEOR score of 56.49% (3). Both systems tuned to translate multiple languages rather than just one pair performed worse than the system only trained for the French-English language pair, showing that even though good results can be achieved with a generalised system, specially trained systems will outperform.

4.2 Primary Research Question 2

Are LLMs the method that will become the prevailing technology in the translation space in future?

The results of this experiment indicate that LLMs cannot quite attain the level of translation accuracy of NMT models – whether they are multilingual or single-language systems. With a small performance difference between the lowest performing NMT model M2M100 (BLEU: 26.03%, METEOR: 56.49%) (3) and the highest performing LLM T5 (BLEU: 24.47%, METEOR: 54.97%) (3) of around 1.5% across both scores. Despite these small differences, the comparison is significant due to the p-value of 0 (16,17) between these systems. When comparing between best-performing NMT system Marian (BLEU: 30.8%, METEOR: 60.59%) (3), and the worst-performing LLM Llama (BLEU: 20.88%, METEOR: 52.6%), there is BLEU gap of nearly 10% and a METEOR gap of nearly 8%. These score differences show that different implementations of LLMs using different approaches can drastically alter translation quality, paving the way for new LLM approaches to be used in the future. Consideration must also be made for LLMs' ability to perform general tasks beyond translation such as text generation, these extra facilities could lead to users taking a small hit in translation quality to have a single solution for all their problems, rather than dedicated systems for each task. However, LLMs incredibly high resource costs for similar or worse translation results limits their ability to spread as a translation tool, as training and running them requires huge time and infrastructure investments. In time, LLMs should become the prevailing technology as customers who already use LLMs will want translation capabilities included. NMT and LLM approaches may also be combined in a similar vein

to how T5 works, allowing for the translation quality of NMT systems, alongside the other abilities of LLMs.

4.3 Secondary Research Question

In the category of LLMs, how do the emergent capabilities of LLMs which have not been trained to do translation tasks compare to the capabilities of LLMs which have explicitly been trained to do translation tasks? Using T5 as the model explicitly trained for translation and Llama as the model with emergent capabilities it is clear there is a significant difference in translation quality between the two. T5's scores (BLEU: 24.47%, METEOR: 54.97%) are higher than Llama's (BLEU: 20.88%, METEOR: 52.6%) with the larger difference in BLEU score of nearly 5% compared to the difference in METEOR score of just over 2%. This gap between translation scores shows specialised training for an LLM significantly enhances translation precision while only slightly enhancing overall translation quality. This suggests that for situations where accuracy of translation is paramount, specifically trained LLMs are a better fit as they will better convey the meaning of the source text.

4.4 MT in Specialised Domains

These results can be extrapolated to gain insights into how these technologies would perform in different situations, such as in specific translation domains like legal or medical disciplines. In these domains translation precision and accuracy are paramount as errors can have serious consequences, as such the systems with the best scores overall, and particularly higher BLEU scores, would fare best in these domains. NMT systems, Marian in particular, are the solution for this given their top overall performance and high BLEU scores, indicating good precision. However, in domains that require less precision and more natural-sounding translations such as creative content like advertising, LLMs could play a key role. If creative companies are already using LLMs for other purposes, their ability to provide good translations that maintain semantic accuracy in an area where precision doesn't matter as much provides these companies with one technical solution for multiple areas.

4.5 Future Developments

As the MT technologies progress, it is important to distinguish which technologies will dominate in the near and long term. In the near term, NMT will remain the dominant technology as its significant performance advantage over other technologies suggests it will be the default choice for the highest-quality translation in the immediate future. In the long term, the best of both LLM and NMT technologies will likely converge, indicated by the small gap between LLM and NMT performance. This idea is also demonstrated by T5's approach of being trained for language translation on top of its general LLM capabilities, incorporating the strengths of both these technologies. As these technologies develop, the trade-off of functionality and computing cost will be prioritised over translation quality as it becomes less of a factor. The large computing costs but extra functionalities of LLMs need to be considered against NMT's lower computing costs but single functionality. Additionally, single-language pair NMT systems will start to be phased out as the close performance gap between Marian and NLLB of less than 1% indicates that multilingual NMT solutions will have equal performance to single-pair solutions.

4.6 Limitations of Analysis

To properly contextualise the analysis made in this section it is important to highlight the limitations of the study that produced the results. The use of automated evaluation metrics without any human evaluation can potentially cause false confidence as there is evidence to show that METEOR and BLEU can miss essential sentiment mistakes in translation (Saadany and Orasan, 2021). In addition, testing on a single high-resource language pair like French-English puts corpus-based translation systems at an advantage as the resources to train them properly, whereas RBMT systems often perform better with low-resource languages (Bayón and Sánchez-Gijón, 2019). The resource constraints in this project could have hindered LLM performance, particularly in the case of Llama, as the precision had to be reduced to 16-bit due to memory constraints and a model with fewer parameters was used. Despite these limitations, the statistical significance of the performance differences shows that the results discussed in this chapter are reliable. These constraints should be considered when interpreting the results of this study and applying its findings.

4.7 Summary

The key findings from this study answer the first research question, definitively showing that NMT is

the superior technology in both semantic accuracy and precision of translation. LLMs closely followed with much lower precision but were closer in semantic accuracy due to their ability to understand the structures of human language with RBMT coming last by a significant amount because of its inability to include semantic context when translating. Despite not being the top-performing technology the results shown by LLMs in this study were very promising, positioning them to become the prevailing technology in the MT field in future, especially when specially trained for translation tasks alongside generative capabilities. Within the LLM field, two different styles of translation were evaluated, emergent translation capabilities and LLMs trained for translation in the form of Llama and T5. T5 had better overall translation quality with a much bigger improvement over Llama in precision, showing that while emergent capabilities are impressive and could be used for non-critical translation, if accurate, precise translation is needed specially trained systems are better. These comparisons can be made with confidence due to the extensive statistical significance testing performed as part of this study, with every p-value being 0 the comparisons between each system are extremely statistically significant and can be evaluated as extremely valid. This study is the first to compare these three translation technologies and as a result, provides unique insight for users or developers considering implementing one of them.

5 Conclusion

The significance of this research is that there is a comprehensive evaluation framework comparing three different MT translation technologies to ensure the accuracy of results and comparison. These translations are also evaluated on both a word-by-word basis and overall semantic basis using multiple evaluation metrics, something many studies lack. The translation task itself covers multiple domains, allowing a true demonstration of each system's more diverse capabilities. The study also implements statistical analysis suggestions by Marie et al. in order to ensure the significance of the findings, leading to confidence that these results can be used to make informed decisions when using these systems in future. The development of a framework like this provides a consistent benchmark future technologies can be measured against. This paper also offers key insights into the current MT space and its potential future trajectory. The results of this study are directly relevant to the automatic extraction of socio-political events in multilingual contexts, where the use of automatic translation methods may be necessary.

5.1 Limitations

Despite this project's successes in creating effective results, multiple resource constraints limited the scope of the research. Computing restraints lead to smaller models being used – particularly when it came to LLMs – with Llama's 3.3 billion parameter model having to be used, despite the availability of larger models. Llama's precision also had to be reduced due to memory constraints with the hardware used. The use of the French-English language pair also favours data-driven approaches as it is a high-resource language pair with plenty of data available to train systems that need it. If this paper were to be repeated with more time allocated more language pairs from different language families would be added to assess the efficacy of each system with different grammatical structures and vocabularies. Statistical models would also be assessed to provide even more context of how different technologies perform.

5.2 Future Work

Future work directly stemming from this research could involve creating both broader and more specific studies. Future research projects with access to more compute or paid APIs can use larger, more performant models such as LLMs with 100 billion or more parameters. This allows better insight into very current technologies in a way that is unavailable with open-source resources. Another avenue of research developed from this would be repeating the same study with more RBMT systems on low-resource languages. This reverses the dynamic of corpus-based systems having an advantage allowing RBMT to show its use in more niche scenarios. A final branch of study resulting from this project would be developing and investigating hybrid NMT-LLM approaches to translation. These would also have to be evaluated from an LLM perspective to ensure the different training method would not affect its generative capabilities. This research would heavily advance the field of MT potentially removing the need for compromise.

Acknowledgements

This work has been partially supported by the CIDEXG/2023/12 project, funded by the Generalitat Valenciana

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

María Do Campo Bayón and Pilar Sánchez-Gijón. 2019. Evaluating machine translation in a low-resource language combination: Spanish-Galician. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 30–35, Dublin, Ireland. European Association for Machine Translation.

Steven Bird, Ewan Klein, and Edward Loper. 2009. Natural language processing with Python: analyzing text with the natural language toolkit.

Antonio M Corbí-Bellot, Mikel L Forcada, Sergio Ortiz-Rojas, Juan Antonio Pérez, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, Iñaki Alegria, and Kepa Sarasola. An Open-Source Shallow-Transfer Machine Translation Engine for the Romance Languages of Spain.

Marta R. Costa-Jussà, Mireia Farrús, José B. Mariño, and José A. R. Fonollosa. 2012. Study and Comparison of Rule-Based and Statistical Catalan-Spanish Machine Translation Systems. *Computing and Informatics*, 31(2):245–270. Number: 2.

D. I. De Silva and D. G. P. Hansadi. 2024. Enhancing machine translation: Cross-approach evaluation and optimization of rbmt, smt, and nmt techniques. In *2024 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, pages 1–8.

Ritesh Kumar Dwivedi, Parma Nand, and Om Pal. 2025. Hybrid NMT model and comparison with existing machine translation approaches. *Multidisciplinary Science Journal*, 7(4):2025146–2025146.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond English-Centric Multilingual Machine Translation. ArXiv:2010.11125 [cs].

Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seo-

hyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 Herd of Models. ArXiv:2407.21783 [cs].

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast Neural Machine Translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–

121, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers. ArXiv:2106.15195 [cs].

Vandan Mujadia, Ashok Urlana, Yash Bhaskar, Penumalla Aditya Pavani, Kukkapalli Shravya, Parameswari Krishnamurthy, and Dipti Misra Sharma. 2023. Assessing Translation capabilities of Large Language Models involving English and Indian Languages. ArXiv:2311.09216 [cs].

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. ArXiv:1910.10683 [cs].

Sreelekha. S and Pushpak Bhattacharyya. 2017. Comparison of smt and rbmt; the requirement of hybridization for marathi-hindi mt.

Hadeel Saadany and Constantin Orasan. 2021. BLEU, METEOR, BERTScore: Evaluation of Metrics Performance in Assessing Critical Translation Errors in Sentiment-oriented Text. In *Proceedings of the Translation and Interpreting Technology Online Conference TRITON 2021*, pages 48–56. ArXiv:2109.14250 [cs].

Fedor Sizov, Cristina España-Bonet, Josef Van Genabith, Roy Xie, and Koel Dutta Chowdhury. 2024. Analysing translation artifacts: A comparative study of LLMs, NMTs, and human translations. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1183–1199, Miami, Florida, USA. Association for Computational Linguistics.

NLLB Team, Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Jeff Wang, and Meta Ai. No Language Left Behind: Scaling Human-Centered Machine Translation.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raul Vazquez, and Sami Virpioja. 2022. Democratizing Neural Machine Translation with OPUS-MT. ArXiv:2212.01936 [cs].

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. ArXiv:2302.13971 [cs].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. ArXiv:1706.03762 [cs].

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Bootstrap distributions and statistics tables

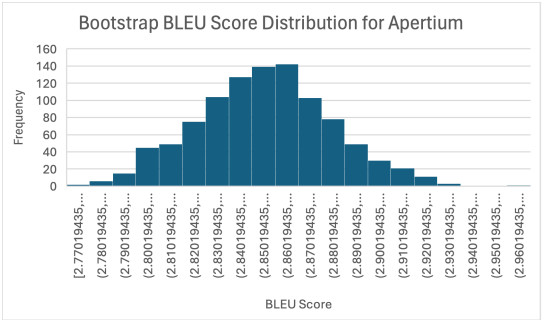

Figure 4: Apertium BLEU score bootstrap distribution

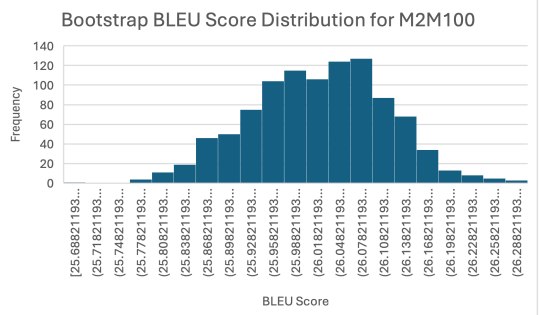

Figure 5: M2M100 BLEU score bootstrap distribution

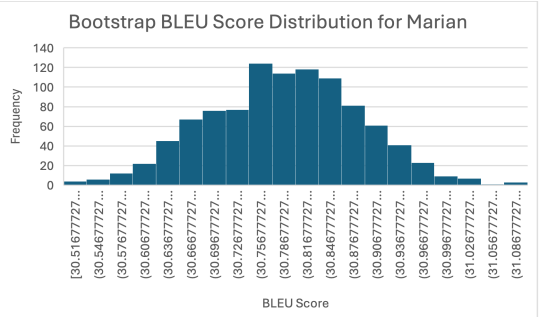

Figure 6: Marian BLEU score bootstrap distribution

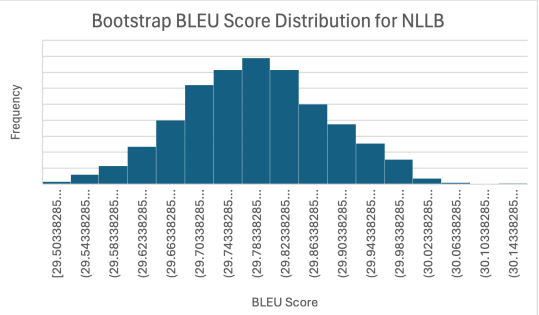

Figure 7: NLLB BLEU score bootstrap distribution

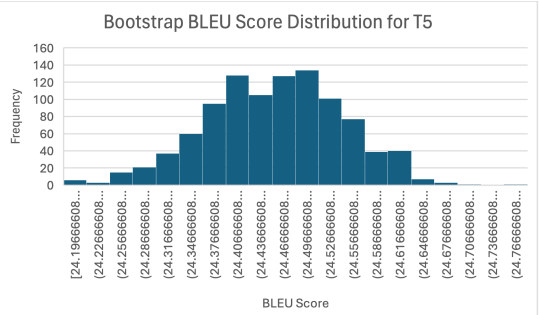

Figure 8: T5 BLEU score bootstrap distribution

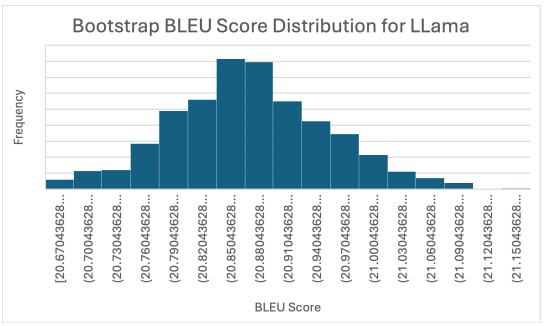

Figure 9: Llama BLEU score bootstrap distribution

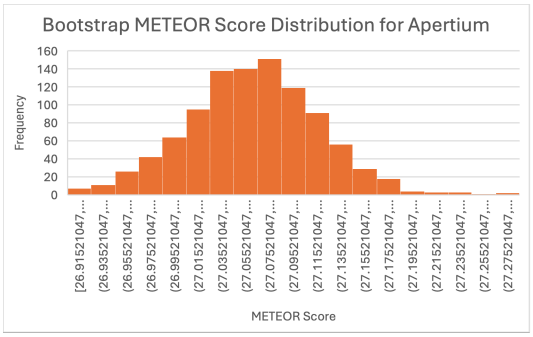

Figure 10: Apertium METEOR score bootstrap distribution

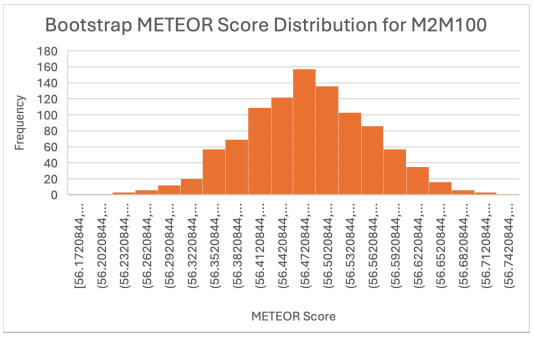

Figure 11: M2M100 METEOR score bootstrap distribution

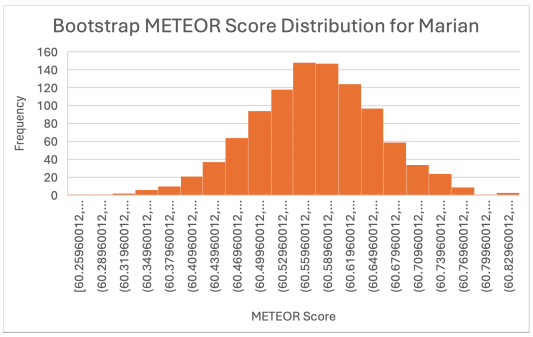

Figure 12: Marian METEOR score bootstrap distribution

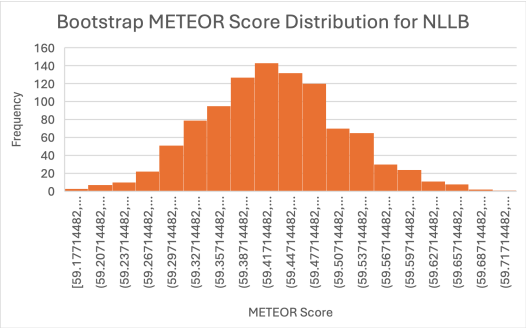

Figure 13: NLLB METEOR score bootstrap distribution

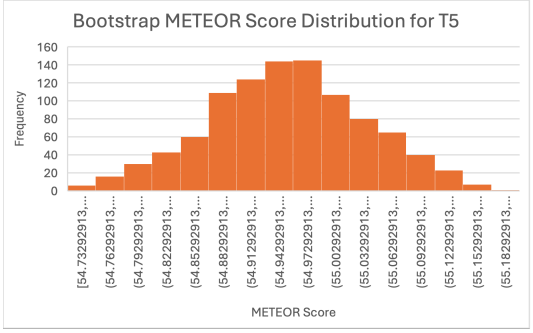

Figure 14: T5 METEOR score bootstrap distribution

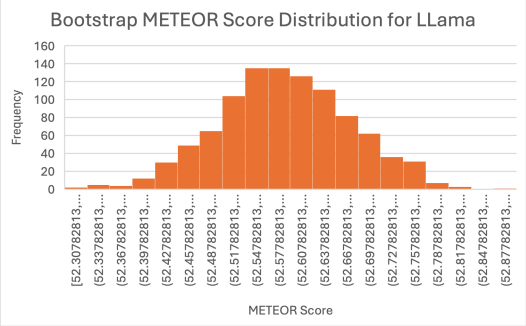

Figure 15: Llama METEOR score bootstrap distribution

Metrics	Llama bleu	M2M100 bleu	Marian bleu	NLLB bleu	T5 bleu	Apertium bleu
mean	20.8855	26.0353	30.8024	29.8023	24.4704	2.8556
std dev	0.0822	0.0919	0.0987	0.1013	0.0888	0.0290
CI	0.0051	0.0057	0.0061	0.0063	0.0055	0.0018
min	20.6704	25.6882	30.5168	29.5034	24.1967	2.7702
max	21.1543	26.2986	31.1079	30.1709	24.7679	2.9626

Figure 16: Statistics table for BLEU bootstrap scores

Metrics	Llama meteor	M2M100 meteor	Marian meteor	NLLB meteor	T5 meteor	Apertium meteor
mean	20.8853	52.6005	60.5876	59.4417	54.9659	27.0714
std dev	0.0823	0.0876	0.0842	0.0873	0.0837	0.0552
CI	0.0051	0.0054	0.0052	0.0054	0.0052	0.0034
min	20.6704	52.3078	60.2596	59.1771	54.7329	26.9152
max	21.1543	52.8834	60.8371	59.7371	55.1862	27.2814

Figure 17: Statistics table for METEOR bootstrap scores

B Link to project github

https://github.com/boydw27/MTInTheAIEra

Steering Towards Fairness: Mitigating Political Bias in LLMs

Afrozah Nadeem, Mark Dras, Usman Naseem
School of Computing, Macquarie University, Australia
afrozah.nadeem@students.mq.edu.au,
{mark.dras,usman.naseem}@mq.edu.au

Abstract

Recent advancements in large language models (LLMs) have enabled their widespread use across diverse real-world applications. However, concerns remain about their tendency to encode and reproduce ideological biases along political and economic dimensions. In this paper, we employ a framework for probing and mitigating such biases in decoder-based LLMs through analysis of internal model representations. Grounded in the Political Compass Test (PCT), this method uses contrastive pairs to extract and compare hidden layer activations from models like Mistral and DeepSeek. We introduce a comprehensive activation extraction pipeline capable of layer-wise analysis across multiple ideological axes, revealing meaningful disparities linked to political framing. Our results show that decoder LLMs systematically encode representational bias across layers, which can be leveraged for effective steering vector-based mitigation. This work provides new insights into how political bias is encoded in LLMs and offers a principled approach to debiasing beyond surface-level output interventions.

1 Introduction

Large Language Models (LLMs) have become foundational tools across a wide spectrum of applications, yet their outputs frequently reflect political and ideological biases, particularly in contexts involving sensitive framing or policy-oriented discourse (Zheng et al., 2023; Afzoon et al., 2025). This problem is particularly pressing in multilingual low-resource settings, where LLMs often produce uneven or culturally misaligned outputs across different languages, amplifying social or political asymmetries (Kumar et al., 2023; Maskey et al., 2025).

Emerging research reveals that a model's ideological leanings are more influenced by input language than by intended sociocultural identity, raising serious concerns about fairness in multilingual settings (Helwe et al., 2025). For instance, the same political statement can elicit starkly different responses when phrased in Urdu versus Punjabi, even within the same model. As illustrated in Figure 1, such biases can result in overconfident responses that reflect tribal or populist framings, potentially skewing downstream interpretations.

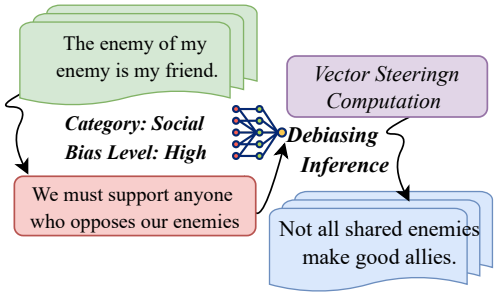

Figure 1: Example of social bias mitigation in our framework. The input PCT statement (4) triggers a high-bias response aligned with tribal loyalty framing.

Prior studies have largely focused on evaluating LLM bias at the output level—either by quantifying stance across Political Compass Test (PCT) statements (Barkhordar et al., 2024) or by cataloging surface-level disparities across languages. However, these approaches stop short of proposing effective and reproducible mitigation strategies that operate within the internal representation space of decoder models (Ejaz et al., 2023).

To address this gap, we investigate a modular activation-based mitigation framework that uses contrastive ideological prompts from the PCT to extract, analyze, and intervene on latent bias directions within decoder LLMs. At the core of the method is the use of Steering Vector Ensembles (SVE): layer-specific vector representations that capture ideological framing and allow for inference-time debiasing without fine-tuning (Siddique et al.,

Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025, pages 52–61, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-006

2025). Our contributions are as follows:

- We present a multilingual bias mitigation method on PCT using Steering Vector Ensembles derived from contrastive political prompts along social and economic axes.
- Our pipeline supports scalable extraction and aggregation of hidden activations across decoder LLMs (e.g., Mistral, DeepSeek) in low-resource languages.
- We demonstrate that ensemble-based interventions reduce bias while maintaining fluency and context relevance, offering a reproducible path toward fairer multilingual LLM behavior.

2 Related Work

2.1 Bias Evaluation via the Political Compass Test (PCT)

The Political Compass Test (PCT) has become a widely used diagnostic tool for probing the political leanings of LLMs (Helwe et al., 2025). Its structured two-axis framework—*economic* (left–right) and *social* (authoritarian–libertarian)—makes it particularly useful for assessing ideological alignment in model responses (Lee et al., 2022).

Early studies (Liu et al., 2024) leveraged the PCT for output-level bias evaluation, prompting models with ideologically framed statements and analyzing completions via stance classification or sentiment scoring. These studies uncovered consistent political leanings in popular LLMs, often skewing toward left-libertarian quadrants (Shen et al., 2023).

Multilingual Political Bias Studies. Recent research has highlighted that language plays a key role in shaping LLM bias. Thapa et al. (2023) translated the PCT into Nepali and found that smaller models exhibited economic-right bias, while larger ones leaned socially left. Nadeem et al. (2025) extended this analysis to low-resource languages (Urdu and Punjabi), showing that models exhibited stronger authoritarian tendencies when generating in low-resource regional languages. Similarly, Helwe et al. (2025) evaluated 15 multilingual LLMs across 50 countries and found that both prompting language and persona assignment significantly influenced model stance—often more so than the nominal national identity (Feng et al., 2023).

These findings collectively underscore that political bias in LLMs is both pervasive and language-conditioned, and that multilingual evaluation is essential to uncovering such disparities. However, all these approaches remain post-hoc, focused solely on surface-level output, and do not probe the internal representation space where ideological bias is likely encoded.

2.2 Steering Vectors and Ensemble Approaches for Mitigation

Beyond evaluation, recent research has explored representation-level mitigation via steering vectors—directional vectors derived from hidden state differences between biased and neutral (or contrastive) inputs. Introduced in contexts like toxicity filtering and sentiment control (Sun et al., 2022), steering vectors operate at the embedding or hidden state level, modifying a model's response without retraining.

More recent work introduced Steering Vector Ensembles (SVE) (Siddique et al., 2025), which aggregate vectors across multiple demographic groups, model layers, or task settings. These ensembles offer improved robustness and generalizability. However, SVE studies have been narrow in scope, often focusing on: *Encoder or encoder-decoder architectures like BERT or T5;*, *Domain-specific settings, such as toxicity or fairness in QA;*, and *English-only applications*, with little attention to ideological framing or multilingual dynamics. Thus, the potential of SVE for open-ended political discourse, particularly in decoder LLMs, remains largely unexplored.

Despite promising advancements, three core gaps remain in the literature:

- **Representation-level blind spots in decoder LLMs:** Most bias studies focus on outputs, leaving open questions about how and where ideological bias is encoded in decoder-only models like Mistral or DeepSeek (Röttger et al., 2024).
- **Lack of systematic contrastive activation pipelines:** There is no open-source or standardized pipeline for extracting contrastive activations (e.g., liberal vs. authoritarian) across layers and prompts in decoder LLMs, particularly for multilingual bias detection.
- **Underutilization of SVE in political contexts:** Steering Vector Ensembles have shown

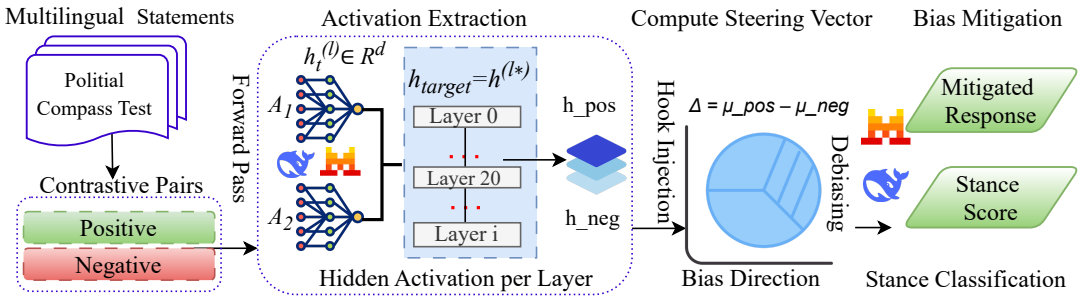

Figure 2: **Bias mitigation pipeline using steering vectors in transformer-based LLMs.** Multilingual statements from the Political Compass Test (PCT) are used to construct contrastive pairs representing opposing ideological stances (positive vs negative). Each pair is passed through a pretrained language model (e.g., DeepSeek-7B), and hidden states $h^{(l)} \in \mathbb{R}^d$ are extracted from each transformer layer. A target layer l^* (e.g., 20) is selected, and its activations are mean-pooled to form h_{pos} and h_{neg}. A bias direction vector is computed as $v = \mu_{\text{pos}} - \mu_{\text{neg}}$, representing the difference between positive and negative class means. This vector is injected via a forward hook into the model's target layer. The modified model then generates a *mitigated response*, which is evaluated using zero-shot stance classification to obtain a final *stance score* (Thapa et al., 2024).

promise in fairness-related domains, but their application to ideological bias mitigation, particularly across languages and political axes, remains under-investigated (Chen et al., 2020).

To address these limitations, our work introduces an activation-based bias mitigation pipeline tailored for decoder LLMs. Our core contributions include: A scalable, multilingual framework for layer-wise activation extraction using PCT-based contrastive pairs; A representation-level analysis method that identifies and aggregates ideological bias directions. And the first integration of Steering Vector Ensembles into decoder-based LLMs for mitigating political bias across both social and economic axes. By bridging output-based evaluation and internal mitigation strategies, we provide a new foundation for probing and correcting ideological bias in multilingual generative models. Although we focus on Political Compass Test (PCT) prompts, our pipeline is modular and could be extended to other domains such as healthcare or education. The ensemble design also improves robustness to prompt framing by aggregating across multiple paraphrases. Our implementation is publicly available to support reproducibility and further work [1].

3 Methodology

This section presents a framework for mitigating political bias in multilingual large language models (LLMs) using contrastive political pairs derived from the Political Compass Test (PCT). The pipeline integrates contrastive pair construction, activation-based analysis, and vector steering. We evaluate steering effectiveness using **Bias Score Reduction** (ΔBias) and response quality measures inspired by the work (Siddique et al., 2025).

3.1 Framework Overview

We introduce a modular pipeline for political debiasing of autoregressive large language models (LLMs) using contrastive prompting and steering vector interventions. Our approach consists of four main stages: *constructing ideologically contrastive prompt pairs* based on translated Political Compass Test (PCT) statements, *extracting hidden activations* from selected transformer layers, *training layer-specific classifiers* to obtain directionally meaningful steering vectors, and *injecting those vectors* during generation to modulate bias.

We implement two steering strategies: *Individual Steering Vectors (ISV)*, where a single vector is derived per layer using logistic regression, and *Steering Vector Ensembles (SVE)*, where vectors from multiple layers are aggregated using response quality-weighted coefficients.

3.2 Multilingual PCT Dataset Preparation

We build on the multilingual PCT dataset proposed by Nadeem et al. (2025), which adapts the 62 standard Political Compass Test (PCT) statements into low-resource languages: Urdu and Punjabi (Smith et al., 2022). We extend this dataset to include English, resulting in six total languages spanning multiple language families (Mostefa et al., 2012).

[1] https://github.com/Afx-Msh/SVE_Mitigation

Each translation was reviewed and verified by regional native speakers, achieving near-perfect inter-annotator agreement with a Fleiss' $\kappa = 0.99$.

The PCT statements cover both ideological axes:

- **Economic axis**: left–right orientation (e.g., redistribution, market policies)
- **Social axis**: libertarian–authoritarian values (e.g., social freedoms, censorship)

Each statement was transformed into a pair of opposing ideological prompts through manual rewriting or structured agreement templates. To ensure semantic divergence and ideological contrast:

1. We computed multilingual sentence embeddings using `sentence-transformers`.
2. Contrastive pairs with cosine similarity below a threshold $\tau = 0.15$ were retained.
3. We limited generation to a maximum of 30 pairs per category, with at most 500 comparisons, to avoid redundancy and infinite pairing loops.

3.3 Target Model and Layer Selection

We selected `deepseek-llm-7b-chat`[2], Mistral model due to its strong multilingual capabilities and transparent architecture. We also evaluate `Mistral-7B-v0.1`[3] to compare bias behavior across model families. We selected layers 8, 12, 16, 20, 24, as layer-wise profiling indicated that mid-level layers encode the strongest ideological signals, whereas early layers capture lexical patterns and very late layers primarily influence fluency.

3.3.1 Individual Steering Vectors (ISV)

We compute a bias-aligned steering vector $\mathbf{v}_l$ for each selected transformer layer l and each ideological axis.

First, we extract hidden activations for positive (e.g., left-leaning) and negative (e.g., right-leaning) prompts. These are standardized using `StandardScaler`, and concatenated to form the input matrix $\mathbf{X} = [\mathbf{A}_{\text{pos}}; \mathbf{A}_{\text{neg}}]$. Corresponding binary labels are assigned as $\mathbf{y} = [1^{n_{\text{pos}}}; 0^{n_{\text{neg}}}]$.

Next, we train a logistic regression classifier with `max_iter=1000` and `random_state=42` to separate the two ideological classes. The resulting classifier weight vector $\boldsymbol{\theta}$ is normalized to unit length to obtain the steering vector $\mathbf{v}_l = \boldsymbol{\theta}/\|\boldsymbol{\theta}\|$.

Finally, to ensure directional consistency, we verify that the expected projection of positive activations exceeds that of negative activations, i.e., $\mathbb{E}[\mathbf{A}_{\text{pos}} \cdot \mathbf{v}_l] > \mathbb{E}[\mathbf{A}_{\text{neg}} \cdot \mathbf{v}_l]$.

3.3.2 Quality Assessment

The vector-quality score for layer l is $q_l = 0.6\,\text{accuracy}_l + 0.4\,\min\left(\frac{\text{separation}_l}{2}, 1.0\right)$. The separation term separation_l measures the normalized effect size between projected activations of opposing ideological prompts: $\text{separation}_l = \frac{|\mu_{\text{pos}} - \mu_{\text{neg}}|}{\text{pooled\_std}}$. The positive and negative projection means are $\mu_{\text{pos}} = \text{mean}(\mathbf{A}_{\text{pos}} \cdot \mathbf{v}_l)$ and $\mu_{\text{neg}} = \text{mean}(\mathbf{A}_{\text{neg}} \cdot \mathbf{v}_l)$.

3.3.3 Steering Vector Ensembles (SVE)

To construct ensemble steering vectors, we aggregate the individual steering vectors (ISVs) computed across the selected layers $l \in \{8, 12, 16, 20, 24\}$ using quality-based weighting. Each vector is assigned a quality score q_l using Equation **??**, and the scores are normalized to obtain weights $w_l = \frac{q_l}{\sum_i q_i}$. The ensemble steering vector is the weighted sum $\mathbf{v}_{\text{SVE}} = \sum_l w_l \mathbf{v}_l$, which is then normalized to unit length $\mathbf{v}_{\text{SVE}} = \frac{\mathbf{v}_{\text{SVE}}}{\|\mathbf{v}_{\text{SVE}}\|}$. SVEs are computed independently for each bias axis (economic and social) and for each language.

3.4 Mitigated Generation via Vector Injection

Bias mitigation is performed by injecting steering vectors into the residual stream of the transformer during generation. Let $h^{(l)}(x)$ denote the last-token hidden activation at layer l for input prompt x. The modified activation is $h^{(l)}(x)' = h^{(l)}(x) + \alpha \mathbf{v}_l$, where α is a tunable steering-strength hyperparameter (default $\alpha = 1.0$) and $\mathbf{v}_l$ is the steering vector. For ISVs, $\mathbf{v}_l$ is injected only into its corresponding layer l, whereas for SVEs the same normalized vector $\mathbf{v}_{\text{SVE}}$ is applied across all selected layers $l \in \{8, 12, 16, 20, 24\}$ simultaneously.

3.5 Bias Detection and Evaluation

We adopt a keyword-based scoring framework to quantify political bias in generated responses. Bias is measured independently along two axes: **social**

[2] https://huggingface.co/deepseek-ai/deepseek-vl-7b-chat

[3] https://huggingface.co/mistralai/Mistral-7B-v0.1

and **economic**. Each axis uses a lexicon of ideologically aligned keywords, adapted for each target language.

Social Bias Lexicons: We adopt a keyword-based scoring framework to quantify political bias in generated responses. Bias is measured independently along two axes: **social** and **economic**. Each axis uses a lexicon of ideologically aligned keywords, adapted for each target language.

Social Bias Lexicons

Progressive: `equality, inclusion, rights, diversity, justice, fair, acceptance`
Conservative: `traditional, family values, moral, heritage, stability, conventional`

Economic Bias Lexicons

Left-leaning: `inequality, exploitation, workers rights, redistribute, regulation, intervention`
Right-leaning: `free market, capitalism, growth, competition, innovation, entrepreneurship`

3.5.1 Bias Score Computation

For a generated response r, we compute the bias score along each ideological axis (social or economic) as $\text{Bias}_{\text{axis}}(r) = \frac{n_{\text{positive}} - n_{\text{negative}}}{n_{\text{total}} + \varepsilon}$, where n_{positive} and n_{negative} are the counts of axis-aligned keywords in the response, $n_{\text{total}} = n_{\text{positive}} + n_{\text{negative}}$, and $\varepsilon = 10^{-8}$ is a small constant to prevent division by zero.

3.5.2 Bias Reduction Metric (ΔBias)

To quantify the effect of steering on bias, we compute the absolute change in bias magnitude before and after mitigation: $\Delta\text{Bias} = \left|\text{Bias}_{\text{original}}\right| - \left|\text{Bias}_{\text{steered}}\right|$. A positive ΔBias indicates successful bias reduction, a negative value suggests overcorrection, and zero indicates no change in bias magnitude.

3.6 Evaluation Protocol

We evaluated each configuration using contrastive pairs per bias axis (social and economic) across languages, comparing outputs with and without steering. Bias reduction was measured using keyword-based and sentiment-based metrics, averaged to compute overall ΔBias. Paired comparisons were used to assess statistical significance across pre- and post-steering outputs.

3.7 Response Quality Metrics

To assess whether debiasing affected output fluency, we compute a combined quality score $Q(r)$ for each response r using a penalty-based formula $Q(r) = \max\big(0,\ \min\big(1,\ 1.0 - P_{\text{length}} - P_{\text{diversity}} - P_{\text{coherence}}\big)\big)$. The quality components are defined as follows. The **length penalty** P_{length} is set to 0.3 if the word count is less than 10, 0.2 if it exceeds 200, and 0.0 otherwise. The **lexical diversity penalty** $P_{\text{diversity}}$ is set to 0.3 if the ratio of unique to total words is less than 0.6, and 0.0 otherwise. The **coherence penalty** $P_{\text{coherence}}$ is set to 0.4 if no grammatically valid sentence is detected using syntactic chunking and dependency parsing.

The final score $Q(r)$ ranges from 0.0 (poor quality) to 1.0 (highly fluent and coherent), allowing for calibrated evaluation of the side-effects of steering-based bias mitigation.

3.8 Stance Score Calculation.

We compute stance scores using a zero-shot classification approach on concatenated Urdu PCT statements and model-generated responses. The classifier is based on `mDeBERTa-v3-base-mnli-xnli`, evaluated against four English labels: *Strongly Agree*, *Agree*, *Disagree*, and *Strongly Disagree*. The model returns confidence scores for each label, which we map to their Urdu equivalents for bilingual interpretability. We then assign numerical scores: ± 10 for strong stances and ± 5 for moderate stances, weighted by their confidence values (Motoki et al., 2024). This process yields a continuous scalar representing both the intensity and direction of the model's political stance in Urdu-language generations.

4 Experimental Environment

All experiments were conducted on GPU-backed RunPod environments to enable scalable and efficient model execution. The hardware included NVIDIA RTX A6000 and A100 GPUs, each with a minimum of 16 GB VRAM, providing sufficient memory for multi-layer activation extraction and vector injection during inference.

4.1 Hyperparameter Configuration

We adopted a consistent generation configuration across all languages and bias axes. The decoding temperature was set to 0.5 to balance lexical diversity with generation consistency. Each response

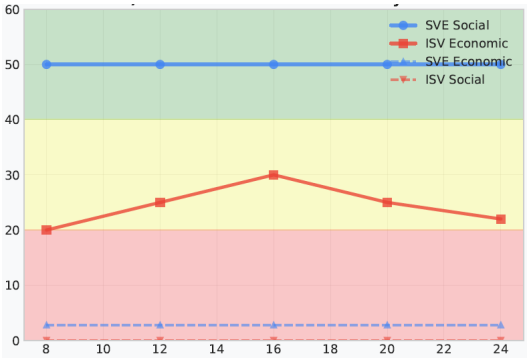

Figure 3: Bias reduction effectiveness across different model layers for SVE and ISV methods.

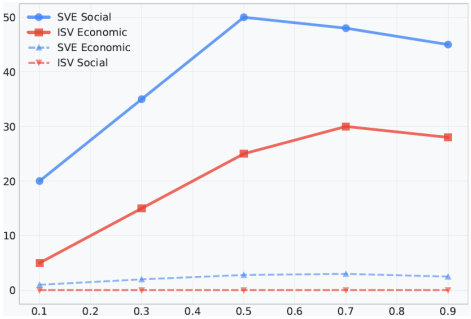

Figure 4: Bias reduction performance of SVE and ISV across Social and Economic dimensions under varying input bias intensities.

was constrained to a maximum of 100 tokens to avoid excessively verbose outputs.

The steering strength was fixed at $\alpha = 1.0$, a value determined through preliminary tuning that offered effective mitigation without degrading fluency. Tokenization employed left-padding, and the end-of-sequence (EOS) token was used as the pad token to maintain decoder compatibility in autoregressive settings.

These settings were held constant throughout all experiments to ensure fair and controlled comparisons between baseline and steered generations.

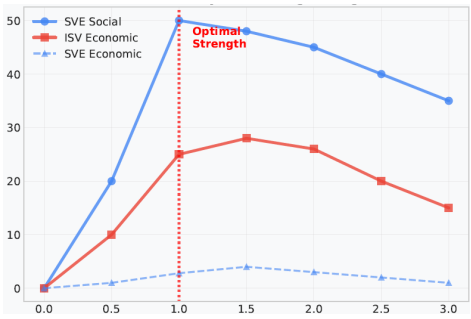

Figure 5: Bias reduction performance of SVE and ISV methods as a function of steering strength.

5 Analysis and Results

5.1 High Resource Language

In our bias mitigation framework, English serves as the high-resource baseline language due to its extensive training data, well-established benchmarks, and consistent performance across models. We use English to construct contrastive political prompts, calibrate steering vectors, and evaluate baseline bias levels before extending the methodology to low-resource languages.

Bias Mitigation Performance. Figure 3 illustrates how bias mitigation effectiveness varies across model layers for both SVE and ISV. SVE for social bias stands out, consistently achieving 50% reduction across all layers and operating in the high-effectiveness zone. ISV for economic bias peaks at layer 16 with 30% reduction but declines. In contrast, SVE for economic bias and ISV for social bias remain below 5% and show little variation. The background shading highlights zones of high (green), moderate (yellow), and low (red) effectiveness, clearly emphasizing the stability and superiority of SVE for mitigating social bias. Figure 4 shows that SVE consistently reduces social bias, while ISV is relatively stronger on economic prompts, highlighting their complementary roles in bias mitigation.

Sensitivity to Steering Strength. Figure 5 illustrates the relationship between steering strength and bias reduction for SVE and ISV methods. A clear optimal point emerges at a steering strength of 1.0, where SVE Social achieves peak effectiveness (50%) and ISV Economic also reaches its maximum (28%). Beyond this threshold, performance gradually declines, indicating that excessive steering may over-correct or destabilize model outputs. SVE Economic, by contrast, exhibits only minimal bias reduction across all strength levels. These results underscore the critical role of hyperparameter tuning—particularly steering strength—in maximizing the effectiveness of vector-based debiasing strategies. Figure 7 shows that SVE excels at reducing social bias in DeepSeek, notably for Traditional Values and Immigration, while ISV is more effective on economic prompts like Taxation. The contrast reveals each method's domain-specific strength, highlighting the benefit of combining them for comprehensive bias mitigation.

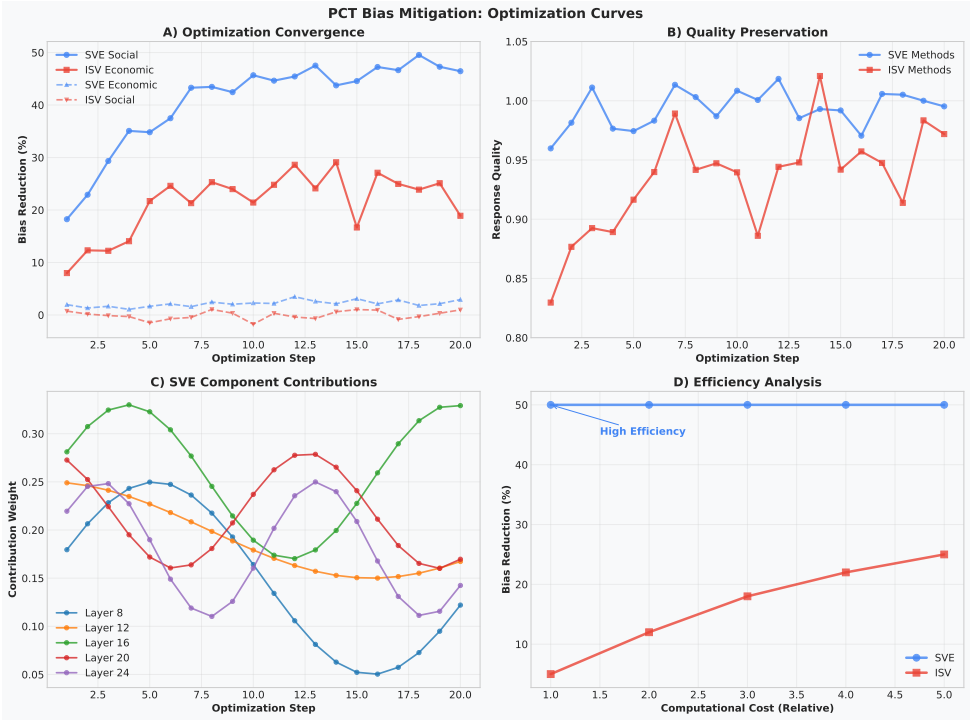

Figure 6: Bias reduction performance of SVE and ISV methods as a function of steering strength.

Model	Econ. (Before)	Soc. (Before)	Econ. (After)	Soc. (After)
Mistral-7B-Instruct-v0.2	2.5	1.23	0.0	0.5
DeepSeek-Chat	-1.0	-1.23	0.0	0.2

Table 1: Bias scores before and after mitigation across models and ideological axes on Urdu language.

Evaluation of Optimization Dynamics. Figure 6 presents a comprehensive comparison of SVE and ISV across key aspects of bias mitigation. SVE for social bias demonstrates steady and effective improvement, achieving up to 50% bias reduction early in the optimization process. It also consistently preserves response quality, maintaining fluency and coherence throughout. In contrast, ISV—particularly for economic bias—shows less stable trends and struggles to match SVE in both fairness and quality. SVE further exhibits adaptability by dynamically leveraging different model layers, particularly mid-layer regions, to optimize its steering effect. Additionally, it delivers strong bias reduction with relatively low computational overhead, making it more cost-efficient than ISV, which requires more resources for smaller gains. These results underscore SVE's advantages in robustness, adaptability, and efficiency. Complementing this, Figure 9 illustrates that ideological distinctions are most pronounced in mid-level layers, aligning with where SVE applies its interventions to guide the model toward more neutral and balanced responses.

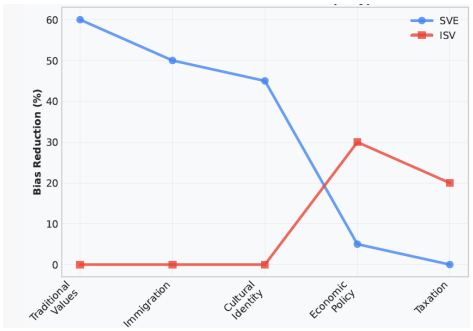

Figure 7: Bias reduction performance on Deep Seek model for SVE and ISV across different contrastive pair types.

The results show in a Table 1 a clear improvement in how both models handle ideological bias after mitigation. Before applying our method, Mistral-7B-Instruct-v0.2 leaned heavily in both economic and social directions, with bias scores as high as 2.5 and 1.23. After mitigation, those scores dropped significantly—closer to neutral—showing that the intervention helped balance its responses.

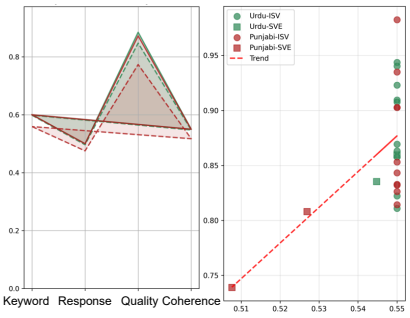

Figure 8: Performance of Bias-Mitigation Methods for Urdu and Punjabi. The left plot compares Keyword Reduction, Response, Quality, and coherence for ISV and SVE.

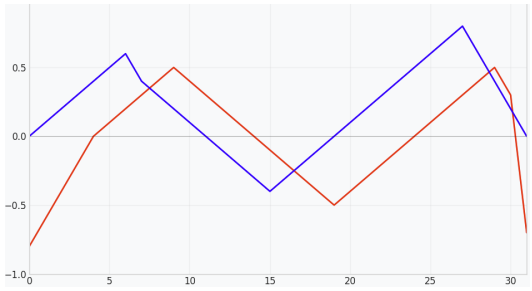

Figure 9: Cosine similarity across hidden layers for contrastive PCT pairs along economic (red) and social (blue) axes.

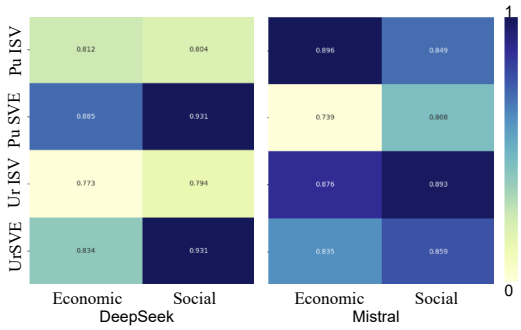

Figure 10: Bias reduction performance of ISV and SVE on model representations across economic and social axes for DeepSeek and Mistral.

Similarly, DeepSeek-Chat started with a noticeable bias in the opposite direction, but also moved toward neutrality after mitigation. Overall, these changes suggest that the approach is effective in steering both models away from extreme positions and helping them generate more balanced, fair outputs. In the Figure 10 how different models respond to bias mitigation methods. In DeepSeek, using SVE noticeably improves the quality of responses, for social topics in both Urdu (Ur) and Punjabi (Pu). The model becomes more fluent and balanced without losing clarity. In some cases like Punjabi economic prompts, SVE lowers the quality. This highlights that not all models benefit from the same debiasing strategy, and choosing the right method depends on the model and language involved.

The results show in the Figure 8 that both the Urdu and Punjabi models handle bias reduction well while keeping their answers natural. They cut out biased keywords to a moderate degree about 0.6 on the scale without over filtering. Response quality stays high, around 0.85 - 0.9, and the overall flow of the replies (coherence) remains steady for both languages. The scatter plot on the right makes it clear that when the overall debiasing score goes up, the quality of the responses also rises, meaning stronger bias mitigation doesn't hurt the readability or sense of the output.

The SVE is effective than ISV for mitigating political bias in decoder-based LLMs. SVE achieved up to 60% reduction on socially framed prompts while preserving response quality, whereas ISV showed moderate gains on economic prompts but was largely ineffective socially. An ablation of ISV, cosine filtering, and ensemble weighting indicates ensembles drive most bias reduction, with ISVs adding targeted improvements on economic prompts. Layer-wise analysis showed mid-level layers carry the strongest ideological signals, and SVE's quality-weighted aggregation across these layers improved robustness and generalization. Both methods performed best at steering strength. DeepSeek benefited most from SVE, producing neutral, fluent responses across Urdu and Punjabi, while Mistral-7B aligned slightly better with ISV on economic axes but lost quality with SVE. We observed degraded quality in Punjabi economic prompts, likely due to vocabulary sparsity, suggesting the value of language-specific calibration in low-resource settings.

6 Conclusion

This study proposes a practical method to reduce political bias in LMs using contrastive prompts from the PCT. SVE outperforms compared to ISV, particularly on socially framed prompts, while preserving response quality. By targeting mid-layer activations with adjustable steering strength, the approach remains efficient and adaptable across models and languages, providing a foundation for fairer multilingual language models.

Limitations

Our approach has several limitations. First, reliance on Political Compass Test (PCT) statements constrains generalizability; although the pipeline is modular and can be applied to other domains (e.g., healthcare, education, gender, race), this remains to be tested. Second, the steering strength parameter (α) and layer selection were manually tuned, limiting adaptability across models; automated calibration could improve robustness. Third, modifying only the last-token activation may not sufficiently propagate steering in longer generations, suggesting a need for dynamic or dialogue-aware steering. Fourth, evaluation relies on keyword-based lexicons, which may miss subtle discursive bias; while stance classification was included, human and discourse-level evaluations are needed. Finally, challenges arose in low-resource settings; Punjabi economic prompts showed reduced quality due to sparse vocabulary, and some entanglement between social and economic axes was observed. Future work could explore language-specific calibration, multi-axis steering, and stance-conditional methods to balance neutrality with context-appropriate stances.

Ethical Considerations

While our approach aims to mitigate political bias in multilingual language models, it raises important ethical concerns. Steering vectors may unintentionally suppress legitimate ideological perspectives or homogenize culturally diverse viewpoints, particularly in low-resource languages. Care must be taken to avoid over-correction, which could result in censorship or erasure of minority opinions. Additionally, the reliance on manually curated keywords and embeddings introduces human biases into the mitigation process. Transparency, documentation, and stakeholder inclusion are essential when deploying such systems. We emphasize that bias mitigation should complement—not replace—broader fairness strategies grounded in cultural, social, and linguistic inclusivity.

References

Saleh Afzoon, Amin Beheshti, Nabi Rezvani, Farshad Khunjush, Usman Naseem, John McMahon, Zahra Fathollahi, Mahdieh Labani, Wathiq Mansoor, and Xuyun Zhang. 2025. Exbigbang: A dynamic approach for explainable persona classification through contextualized hybrid transformer analysis. *arXiv preprint arXiv:2508.15364*.

Ehsan Barkhordar, Surendrabikram Thapa, Ashwarya Maratha, and Usman Naseem. 2024. Why the Unexpected? Dissecting the Political and Economic Bias in Persian Small and Large Language Models. *ELRA Language Resource Association*, pages 410–420. CC BY-NC 4.0.

Wei-Fan Chen, Khalid Al Khatib, Henning Wachsmuth, and Benno Stein. 2020. Analyzing Political Bias and Unfairness in News Articles at Different Levels of Granularity. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 149–154, Online. Association for Computational Linguistics.

Waqas Ejaz, Muhammad Ittefaq, and Sadia Jamil. 2023. Politics triumphs: a topic modeling approach of analyzing news media coverage of climate change in Pakistan. *JCOM*, 22(01):A02.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. *Association for Computational Linguistics*.

Chadi Helwe, Oana Balalau, and Davide Ceolin. 2025. Navigating the political compass: Evaluating multilingual llms across languages and nationalities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17179–17204, Bangkok, Thailand. Association for Computational Linguistics.

Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2023. Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3299–3321, Dubrovnik, Croatia. Association for Computational Linguistics.

Nayeon Lee, Yejin Bang, Tiezheng Yu, Andrea Madotto, and Pascale Fung. 2022. NeuS: Neutral Multi-News Summarization for Mitigating Framing Bias. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3131–3148, Seattle, United States. Association for Computational Linguistics.

Wenhao Liu, Xiaohua Wang, Muling Wu, Tianlong Li, Changze Lv, Zixuan Ling, Jianhao Zhu, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. 2024. Aligning Large Language Models with Human Preferences through Representation Engineering. ArXiv:2312.15997 [cs].

Utsav Maskey, Sumit Yadav, Mark Dras, and Usman Naseem. 2025. Safeconstellations: Steering llm safety to reduce over-refusals through task-specific trajectory. *arXiv preprint arXiv:2508.11290*.

Djamel Mostefa, Khalid Choukri, Sylvie Brunessaux, and Karim Boudahmane. 2012. New language resources for the Pashto language.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: measuring ChatGPT political bias. *Public Choice*, 198(1-2):3–23.

Afrozah Nadeem, Mark Dras, and Usman Naseem. 2025. Framing political bias in multilingual llms across pakistani languages. *arXiv preprint arXiv:2506.00068*.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large Language Model Alignment: A Survey. ArXiv:2309.15025 [cs].

Zara Siddique, Irtaza Khalid, Liam D. Turner, and Luis Espinosa-Anke. 2025. Shifting Perspectives: Steering Vector Ensembles for Robust Bias Mitigation in LLMs. ArXiv:2503.05371 [cs].

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "I'm sorry to hear that": Finding New Biases in Language Models with a Holistic Descriptor Dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tianxiang Sun, Junliang He, Xipeng Qiu, and Xuanjing Huang. 2022. BERTScore is Unfair: On Social Bias in Language Model-Based Metrics for Text Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3726–3739, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Surendrabikram Thapa, Ashwarya Maratha, Khan Md Hasib, Mehwish Nasim, and Usman Naseem. 2023. Assessing Political Inclination of Bangla Language Models. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 62–71, Singapore. Association for Computational Linguistics.

Surendrabikram Thapa, Kritesh Rauniyar, Ehsan Barkhordar, Hariram Veeramani, and Usman Naseem. 2024. Which Side Are You On? Investigating Politico-Economic Bias in Nepali Language Models. *Proceedings of the 22nd Annual Workshop of the Australasian Language Technology Association (ALTA 2024)*, pages 104–117.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. ArXiv:2306.05685 [cs].

wangkongqiang@CASE 2025: Detection and Classifying Language and Targets of Hate Speech using Auxiliary Text Supervised Learning

Kongqiang Wang
School of Information Science,
Yunnan University,
Yunnan Baiyao Street, Kunming,
650500, Yunnan, China.
wangkongqiang@stu.ynu.edu.cn

Peng Zhang
School of Information Science,
Yunnan University,
Yunnan Baiyao Street, Kunming,
650500, Yunnan, China.
zpp1219@gmail.com

Abstract

Our team was interested in content classification and labeling from multimodal detection of Hate speech, Humor, and Stance in marginalized socio-political movement discourse. We joined the task: Subtask A-Detection of Hate Speech and Subtask B-Classifying the Targets of Hate Speech. In this two task, our goal is to assign a content classification label to multimodal Hate Speech. Detection of Hate Speech: The aim is to detect the presence of hate speech in the images. The dataset for this task will have binary labels: No Hate and Hate. Classifying the Targets of Hate Speech: Given that an image is hateful, the goal here is to identify the targets of hate speech. The dataset here will have four labels: Undirected, Individual, Community, and Organization. Our group used a supervised learning method and a text prediction model. The best result on the test set for Subtask-A and Subtask-B were F1 score of 0.6209 and 0.3453, ranking twentieth and thirteenth among all teams.

1 Introduction

First of all, let's introduce the Overview of Shared Task on Multimodal Hate, Humor, and Stance Detection in Marginalized Movement@CASE2025 (Hürriyetoğlu et al., 2025). The complexity of text-embedded images presents a formidable challenge in ML given the need for multimodal understanding of multiple aspects of expression conveyed by them. Particularly, the marginalized movement stands as a prominent subject of online discourse, where text-embedded images like memes serve as vehicles of both solidarity and resistance, reflecting the multifaceted dynamics of attitudes and perceptions within the community and beyond. In this context, the distinction between humor and harm becomes blurred, as memes straddle the line between satire and offense, challenging researchers and platforms alike to navigate the complexities of online content moderation. As one label generally fails to encompass multiple aspects of linguistics, this shared task classifies images on four aspects: hate, targets of hate, stance, and humor as subtasks.

Our group mainly participated in the following two sub-tasks. Subtask A-Detection of Hate Speech: The aim is to detect the presence of hate speech in the images. The dataset for this task will have binary labels: No Hate and Hate. Subtask B-Classifying the Targets of Hate Speech: Given that an image is hateful, the goal here is to identify the targets of hate speech. The dataset here will have four labels: Undirected, Individual, Community, and Organization. We have made tremendous progress in these two sub-tasks.

2 Dataset

In this section, we describe various aspects of task dataset including data collection, annotation guidelines, and dataset statistics. Task dataset comprises 5,063 text-embedded images that encompass memes, posters, and infographics relevant to the LGBTQ+ movement. Official dataset only include images from 2020-2024 as this period saw an upsurge of social media content in this domain (Oz et al., 2023). This also allows task dataset to represent contemporary social media interactions through memes. Note that by the term LGBTQ+, official dataset refer to all gender identities and sexual orientations inclusively.

2.1 Data Collection

To maintain diversity in the dataset, organizer collected data from three popular social media platforms: Facebook, Twitter, and Reddit, through manual search and extraction. For Twitter, organizer used hash tags such as #lgbt, #pride, #trans, #transrights, #nonbinary, and #genderidentity to filter images related to LGBTQ+ discussions. For Facebook, organizer targeted groups that frequently

Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025, pages 62–70, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-007

discussed LGBTQ+ content. Similarly, for Reddit, organizer identified subreddits where discussion related to LGBTQ+ was more prominent. Further, to ensure the relevance and quality of the dataset, the data collection process was subject to filtering criteria. Detailed filtering criteria for our dataset can be found in (Shah et al., 2024). As different annotators may encounter and collect the same image, organizer sequentially employed two image deduplication tools: dupeGuru[1] and difPy[2], to search for duplicates and retain the highest quality image out of each batch of duplicates. Organizer used the OCR application provided by Google Cloud Vision API[3] to extract textual data from the images. Organizer removed non-alphanumeric elements such as special characters, hyperlinks, symbols, and non-English characters to reduce noisy text data and ensure data quality. Note that the text may occasionally contain unintentional noisy artifacts.

3 Data Annotation

Organizer engaged five experienced annotators, well-versed in NLP and computational linguistics, to annotate data samples for PrideMM. The annotators had a prior understanding of the LGBTQ+ movement and meme archetypes on social media. Organizer presented them with comprehensive annotation guidelines to ensure uniform and unbiased annotations, and asked them to annotate each image separately for all four tasks. A 3-phase annotation schema was used to ensure accurate and consistent annotations. First, a dry run was conducted to evaluate the understanding of the annotation guidelines among the annotators where every annotator was given an identical batch of 50 images for annotation. Second, a revision phase was conducted where every annotator was given another identical batch of 200 images and received a revised set of instructions based on the results of the first phase. Finally, in the consolidation phase, the annotators annotated a final batch of 50 images while discussing and revising the annotation guidelines until a consensus was reached. These steps were taken to minimize misannotations and noisy labels in the PrideMM dataset. The meticulously devised annotation guidelines were followed to ensure consistency in the annotations. Each image in their dataset was independently annotated for the three aspects and one sub-class, apart from the connection between 'Hate' and 'Hate Targets'.

[1]https://github.com/arsenetar/dupeguru

[2]https://github.com/elisemercury/Duplicate-Image -Finder

[3]https://cloud.google.com/vision/docs

4 Annotation Guidelines

In this section, organizer describe the annotation guidelines used to annotate the dataset. They devise separate guidelines for each of the four tasks.

Hate Speech. This task aimed to identify instances of hate speech in the images. The primary focus was on identifying images that intentionally conveyed hateful sentiments. Annotators needed to distinguish between images expressing strong disagreement without resorting to offensive language and those containing genuine elements of hate speech. This differentiation aimed to guarantee accurate labeling, ensuring that images conveying genuinely hateful sentiment through visual content, language, or a combination of both were appropriately identified.

Hate Targets. This task required annotators to identify the targets of hate in hateful images by classifying the images into one of the four classes: Undirected, Individual, Community, and Organization. Images were labeled as Undirected when they targeted abstract topics, societal themes, or ambiguous targets like 'you' that were not directed toward any specific individuals, entities, or groups. Hateful images targeting specific people including political leaders, celebrities, or activists like 'Joe Biden' and 'J.K. Rowling' were annotated as Individual. Likewise, the label Community was used for instances of images targeting broader social, ethnic, or cultural groups like 'LGBT' or 'trans'. Lastly, images targeting corporate entities, institutions, or similar organizations like 'Chick-fil-A' and 'government" were annotated as Organization.

Stance. This task involved annotating the images into either of three distinct categories: Support, Oppose, and Neutral, determined by their stance within the context of the LGBTQ+ movement. The Support label was given to images that expressed support towards the goals of the movement, agreed with efforts in fostering equal rights for LGBTQ+ individuals, and promoted awareness for the movement's goals. The Oppose label was given to images that conveyed disagreement with the goals of the movement, denied the problems faced by individuals who identified as LGBTQ+, and dismissed the need for equal rights and acceptance. The Neutral label was given to images that were contextually relevant to the movement but

did not exhibit support or opposition towards the movement.

Humor. In this task, annotators were asked to identify images showcasing humor, sarcasm, or satire related to the LGBTQ+ Pride movement. Annotators were instructed to discern the presence of humor in the images regardless of whether they presented a lighthearted or insensitive perspective on serious subjects. Note that annotators were asked to annotate images based on whether the creator of the image intended for it to be humorous, and not based on whether the annotator personally found it humorous. This task aimed to capture the nuanced use of text-embedded images for comedic or satirical purposes, thereby helping disentangle hate and humor in the images related to this movement.

5 Statistics and Inter-Annotator Agreement

Table 1: Dataset Statistics for PrideMM. The data consists of 5,063 samples for Hate, Stance, and Humor tasks, and 4,482 samples for the Target classification task.

Task	Label	#Samples	%
Hate	No Hate	2,581	50.97%
	Hate	2,482	49.03%
Target	Undirected	771	31.07%
	Individual	249	10.03%
	Community	1,164	46.90%
	Organization	298	12.00%
Stance	Neutral	1,458	28.80%
	Support	1,909	37.70%
	Oppose	1,696	33.50%
Humor	No Humor	1,642	32.43%
	Humor	3,421	67.57%

Table 1 shows the distribution of images in PrideMM across all class labels. For the hate detection task, the dataset has a balanced distribution of binary labels. The target classification task exhibits a heavily imbalanced distribution. Given the context of this study, most hateful images convey undirected hate or are targeted toward communities, with a low frequency of hate against individuals and organizations. For the stance classification task, the number of images is well-balanced across three labels. On the other hand, as memes are often meant to be humorous, the majority of the images in the dataset are annotated to humor.

Organizer used the Fleiss' Kappa (κ) (Falotico and Quatto, 2015) as a statistical measure to assess the inter-annotator agreement across all four tasks. For Task A (Hate Speech detection), κ was 0.66/0.74 in the dry run and final phase respectively, for Task B(Target detection), κ was 0.68/0.81, for Task C (Stance detection), κ was 0.62/0.75, and for Task D (Humor detection), κ was 0.60/0.74. The increase in κ from the dry run phase to the final phase across all tasks reflects the effectiveness of the 3-phase annotation schema.

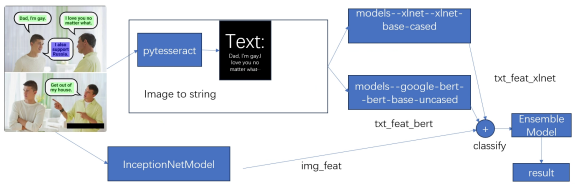

Figure 1: The framework diagram of the Ensemble model.

6 Methodology

6.1 Ensemble Model

MultiModal Hate Speech Detection In our task, there are two modalities of image and text. We use the pytesseract tool to extract the text in the image and obtain the features of the text through the xlnet and bert models. The corresponding image features of the pictures are obtained through the inception net model, and then the classification results are finally obtained through the classifier of the ensemble model. The framework diagram of the model is shown in the Figure 1. The classifier in the Ensemble model uses a neural network classifier. The specific parameters of the classifier are shown in the Table 2.

Table 2: Neural Network classifier structure for Subtask A.

layer number	layer type	Input dimension	Output dimension / parameter
1	Linear	image_features + xlnet_hidden + bert_hidden	512
2	ReLU	512	512
3	BatchNorm1d	512	512
4	Dropout	512	512 (p=0.3)
5	Linear	512	256
6	ReLU	256	256
7	BatchNorm1d	256	256
8	Dropout	256	256 (p=0.3)
9	Linear	256	num_classes = 2

MultiModal Hate Targets Classification The method of MultiModal Hate Targets Classification is the same as that of MultiModal Hate Speech Detection, except that we use two different classifiers for experiments. It is found that using the Linear classifier alone is slightly better than using

the Neural Network classifier. Their structures are respectively shown in Table 3 and Table 4.

Table 3: Neural Network classifier structure for Subtask B.

layer number	layer type	Input dimension	Output dimension/parameter
1	Linear	image_features + xlnet_hidden + bert_hidden	512
2	ReLU	512	512
3	BatchNorm1d	512	512
4	Dropout	512	512 (p=0.3)
5	Linear	512	256
6	ReLU	256	256
7	BatchNorm1d	256	256
8	Dropout	256	256 (p=0.3)
9	Linear	256	num_classes = 4

Table 4: Linear classifier structure.

layer number	layer type	Input dimension	Output dimension/parameter
1	Linear	image_features + xlnet_hidden + bert_hidden	num_classes = 4

6.2 K-max pooling neural network with recurrent learning rate (CLR) scheduling

In the Ensemble Model, the addition of image features leads to poor classification effect of the model. Relying solely on the text extracted from the images may improve the performance of the model. So in this model, we first use the Google Vision API to extract text from images and classify the text content. Facts have also proved the usefulness of our conjecture.

6.2.1 Problem Setup

We address the binary and multi-task classification problem on textual content, where each sample may be annotated with hierarchical labels: a binary label for Task 1 (e.g., Hate vs. Not-Hate) and a fine-grained four-class label for Task 2 (e.g., Undirected , Individual , Community , Organization). The dataset is preprocessed and split into training and test sets accordingly, using a k-fold cross-validation scheme to improve generalizability and model robustness.

6.2.2 Preprocessing and Tokenization

All text inputs are tokenized using nltk.RegexpTokenizer, preserving only word characters. A Keras Tokenizer is then applied to convert the text into sequences of word indices. These sequences are padded to a maximum sequence length based on the longest sample in the training set.

6.2.3 Embedding Layer Construction

We utilize pre-trained GloVe embeddings (840B.300d) to initialize the word embedding matrix. Each word in the vocabulary is mapped to a 300-dimensional dense vector. If a word is missing from the GloVe vocabulary, it is assigned a vector initialized from a normal distribution with the same mean and standard deviation as the pre-trained embeddings.

6.2.4 Model Architecture

The core of the proposed model is based on a K-max pooling neural network architecture enhanced by embedding initialization and dense transformations:

- Input Layer: Tokenized and padded sequences.
- Embedding Layer: Initialized with pre-trained GloVe vectors, this layer is frozen (non-trainable) during training.
- K-Max Pooling Layer: Extracts the top-k activations across the sequence dimension, effectively capturing the most informative word features regardless of position.
- Dense Transformation: The pooled features are passed through a fully connected layer with tanh activation, followed by a softmax classification head.

This simple yet effective architecture is chosen to reduce overfitting and training time while maintaining competitive performance.

6.2.5 Learning Rate Scheduling

To improve convergence, we apply Cyclic Learning Rate (CLR) scheduling, as proposed by (Smith, 2017). Specifically, we use the exp_range policy to periodically vary the learning rate between 0.001 and 0.006 using a base-2 exponential decay factor (gamma=0.99994). This prevents premature convergence and encourages the model to escape local minima during training.

6.2.6 Training Strategy

The model is trained using the Adam optimizer with default hyperparameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 1e-8$). The loss function used is categorical_crossentropy, and additional evaluation metrics include accuracy and a custom-defined F1-score metric implemented using Keras backend operations.

To ensure reliable evaluation, we perform 4-fold cross-validation. In each fold:

- The model is trained on k-1 folds and validated on the remaining fold.

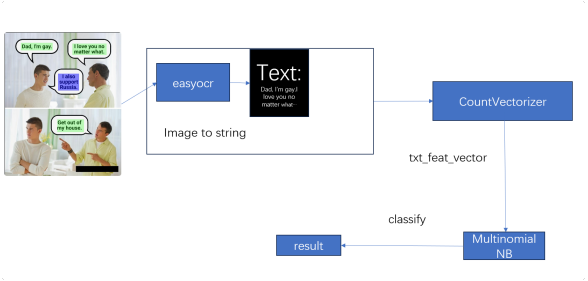

Figure 2: The framework diagram of the Multinomial Naive Bayes classification model.

- Predictions on the test set are collected and averaged across folds for final ensemble predictions.

6.2.7 Prediction and Submission

After training, the ensemble of models predicts the labels for both tasks. The final predictions are computed by averaging softmax outputs across all folds and selecting the label with the highest averaged probability. Results are saved into submission files corresponding to each task.

6.2.8 Evaluation

The model performance is evaluated using macro-averaged precision, recall, and F1-score for both tasks. This metric choice reflects the need to account for class imbalance in the multi-class setting.

6.2.9 Summary

Overall, our model integrates pre-trained embeddings, efficient pooling mechanisms, cyclic learning rates, and ensemble training via cross-validation to deliver a robust solution for binary and multi-task text classification.

6.3 Multinomial Naive Bayes classification model applying easyocr text extraction

We use the English (en) version in the easyocr application for text extraction, and then perform text vectorization through the CountVectorizer in the sklearn.feature_extraction.text library. Finally, the Multinomial Naive Bayes classification model is used for classification. The overall architecture diagram of the model is shown in the Figure 2.

6.3.1 Introduction to EasyOCR

EasyOCR is an open-source, deep learning-based Optical Character Recognition (OCR) tool developed by Jaided AI, supporting text recognition in over 80 languages. This library is implemented based on PyTorch, adopts multi-layer convolutional neural network (CNN) and sequence modeling (such as LSTM) structures, and combines CTC (Connectionist Temporal Classification) loss for the training and recognition of unaligned text sequences.

The main features of EasyOCR are as follows: Multilingual support. Built-in support for multiple languages including Chinese, English, Japanese, Korean, etc., suitable for multilingual text scenarios; Strong end-to-end recognition capability. It can automatically detect text areas from the original image and recognize their contents; No complex preprocessing required. Supports complex backgrounds, slanted text and multi-font recognition; Simple and easy to use. The API interface is friendly and suitable for quick integration into applications; Strong scalability. Users can customize training data and fine-tune the model to adapt to character sets or styles in specific domains.

Brief description of the workflow:

- Text Detection: The CRAFT (Character Region Awareness for Text Detection) model is adopted to locate the text regions in the image;
- Text Recognition: Use deep neural networks (CNN + LSTM + CTC) to conduct sequence modeling and character recognition for the extracted text regions;
- Language post-processing (optional) : Language-level correction in combination with dictionaries or rules.

6.3.2 Principle Explanation of CountVectorizer

CountVectorizer is one of the most fundamental and commonly used methods in text feature extraction, used to convert text data into vector format, facilitating subsequent processing by machine learning models. Its basic idea is: to count the number of occurrences of each word (or n-gram) in the text and take these count values as the elements of the feature vector.

Fundamental: The main processing flow of CountVectorizer is as follows:

- Tokenization. Segment the text to divide sentences or documents into individual words (tokens).
- Vocabulary Building. Based on all the input text, count all the words that have appeared

(or the specified n-gram), and assign a unique index to each word.

- Word Frequency Statistics (Vectorization). For each input text, count the number of times each word appears in the vocabulary to form a sparse vector.

Suppose the vocabulary is ['apple ', 'banana', 'orange'] and the text is 'apple orange orange', then the conversion result is [1, 0, 2].

Mathematical Treatment: Given a example: Vocabulary size V; Input text dataset $D = \{d_1, d_2, \ldots, d_n\}$; Each text d_i is converted to a vector $\vec{x_i} \in R^V$ Then:

$$x_{ij} = \text{count}(\text{term}_j \text{ in } d_i) \tag{1}$$

where x_{ij} represents the occurrence frequency of the j-th term in the vocabulary within the i-th document.

Configurable Parameters: CountVectorizer provides several key parameters that affect the vectorization:

- `ngram_range=(1,1)`: Specifies the use of 1-gram (single words), 2-gram (word pairs), etc.
- `stop_words='english'`: Removes English stop words
- `max_features=1000`: Retains only the top 1000 most frequent terms
- `min_df` / `max_df`: Filters terms that appear too rarely or too frequently
- `binary=True`: Does not count frequencies, only marks whether terms appear or not

Summary: CountVectorizer is a simple and efficient feature extraction method for text frequency; It produces a one-hot sparse matrix, suitable for text classification and retrieval modeling; It does not consider word order or context meaning, but performs well in many tasks that depend on word frequency; For more complex language modeling, it is often combined with TF-IDF, word vectors (Word2Vec), or Transformer models.

6.3.3 Explanation of the principle of Multinomial Naive Bayes

Basic Conception: Multinomial Naive Bayes (Polynomial Naive Bayes) is a type of Naive Bayes classifier, mainly used for classification problems of discrete features, especially suitable for text classification, such as spam recognition, sentiment analysis, news classification, etc. It is based on Bayes' theorem and assumes that features are conditionally independent of each other, that is, the source of "naive".

Core Formula: Given a document d, we want to find the most likely class $y \in \{c_1, c_2, \ldots, c_k\}$. Using Bayes' theorem:

$$P(y|d) = \frac{P(d|y) \cdot P(y)}{P(d)} \tag{2}$$

Since $P(d)$ is the same for all classes, we only need to maximize the numerator:

$$\hat{y} = \arg\max_y P(y) \cdot P(d|y) \tag{3}$$

In the multinomial model, the document is represented as a word frequency vector $\vec{x} = (x_1, x_2, \ldots, x_n)$, where x_i is the frequency of word w_i in the document. Assuming word independence, the likelihood of document under class y is:

$$P(d|y) = \prod_{i=1}^{n} P(w_i|y)^{x_i} \tag{4}$$

The final classification formula:

$$\hat{y} = \arg\max_y \log P(y) + \sum_{i=1}^{n} x_i \cdot \log P(w_i|y) \tag{5}$$

Parameter Estimation: Prior probability $P(y)$: The occurrence ratio of class in training data.

$$P(y = c) = \frac{\text{count}(y = c)}{\text{total samples}} \tag{6}$$

Conditional probability $P(w_i|y)$: The relative frequency of word w_i in class y. To avoid zero probabilities, Laplace smoothing is applied.

$$P(w_i|y) = \frac{\text{count}(w_i \text{ in } y) + \alpha}{\sum_j \text{count}(w_j \text{ in } y) + \alpha \cdot V} \tag{7}$$

α: Smoothing parameter, usually 1 (Laplace) or a small value (Lidstone). V: Vocabulary size.

Advantages and Applications: Advantages: Simple model with high computational efficiency; Performs well on high-dimensional sparse data (e.g., text); Fast training speed, no need for gradient descent; Not prone to overfitting, good generalization. Application scenarios: Text classification (news, reviews, spam filtering); Discrete counting data (click-through rates, purchase data). Compared with other Naive Bayes variants, see Table 5.

Table 5: Naive Bayes Model Variants and Applications

Model	Feature Type	Application Domain
BernoulliNB	Binary features	Text word presence / absence
MultinomialNB	Count features	Text word frequency
GaussianNB	Continuous features	Images, sensor data, etc.

Conclusion: MultinomialNB is one of the most classic models in text classification; It classifies by counting word frequencies in text; Simple and efficient, making it a good baseline model for text classification; Performs well when combined with `CountVectorizer` or `TfidfVectorizer`.

7 Experimental Results

According to the official instructions (Thapa et al., 2025a) for OCR extraction: If participants want to extract OCR, they can use Google Vision API, tesseract, EasyOCR, etc. In the paper that benchmarks this dataset, organizer have used Google Vision API to extract OCR for training the models. Since a lot of participants may not have access to the vision API, they can use the extracted text from organizer's benchmark paper (Bhandari et al., 2023). So when we used the Google vision API to extract the text version database, we directly used the dataset indicated by the official. For the other two methods, namely the pytesseract and EasyOCR methods, we manually extracted them through our own python script code. The complete code of this entire project can be found at our GitHub address[4].

For Subtask A-Detection of Hate Speech and Subtask B-Classifying the Targets of Hate Speech, the results obtained by our three and four methods on the test set are shown in Table 6 and Table 7.

Table 6: The results obtained by our three methods for Subtask A-Detection of Hate Speech on the test set.

Model	Recall	Precision	F1	Accuracy
Ensemble Model (Neural Network classifier)	0.4928	0.4733	0.3761	0.4852
K-max Pooling Neural Network	0.5925	0.6389	0.5585	0.5976
Multinomial Naive Bayes Classification Model	0.6365	0.6591	0.6209	0.6331

[4]https://github.com/WangKongQiang/Case2025

Table 7: The results obtained by our four methods for Subtask B-Classifying the Targets of Hate Speech on the test set.

Model	Recall	Precision	F1	Accuracy
Ensemble Model (Neural Network classifier)	0.2500	0.1175	0.1598	0.4699
Ensemble Model (Linear classifier)	0.2525	0.2503	0.2477	0.3695
K-max Pooling Neural Network	0.2544	0.2846	0.1723	0.4739
Multinomial Naive Bayes Classification Model	0.3322	0.5552	0.3453	0.4779

8 Discussion

For Multimodal Hate, Humor and Stance Detection in Marginalized Movement@CASE2025 sharing task, we referred to the relevant tasks of CASE 2024 (Thapa et al., 2024) and CASE 2023 (Thapa et al., 2023) shared tasks on multimodal hate speech detection and derived our own method. Although the effect of the experiment needs to be strengthened. However, these contents and ideas have given us a lot of inspiration. Multimodal content analysis is a longstanding tradition of the CASE workshop series. We believe that with our further research and more detailed optimization and training of the model, we will achieve even greater success in future competitions.

9 Conclusion

We employed multiple methods in Subtask A-Detection of Hate Speech and Subtask B-Classifying the Targets of Hate Speech, which respectively involved the transformer model, deep learning models and machine learning models in these two tasks. Our final leaderboards are respectively shown in the Table 8 and in the Table 9.

Table 8: The Final Leaderboard of Subtask A: Detection of Hate Speech.

#	User	Team Name	Recall	Precision	F1	Accuracy
1	wangxiuxian	TUJ-MI	0.8422 (1)	0.8422 (1)	0.8422 (1)	0.8422 (1)
2	Ryuan		0.8288 (2)	0.8291 (2)	0.8284 (2)	0.8284 (2)
3	jiarranDiana	IMU-L	0.8211 (3)	0.8217 (3)	0.8205 (3)	0.8205 (3)
4	ray-sushant	Phantom Troupe	0.8189 (4)	0.8193 (4)	0.8191 (4)	0.8190 (4)
5	Neuron-Force	MemeMasters	0.8086 (5)	0.8086 (5)	0.8086 (5)	0.8086 (5)
6	shrutigurung	Multimodal Kathmandu	0.8004 (6)	0.8028 (6)	0.8005 (6)	0.8008 (6)
7	Sujal_Maharjan		0.7932 (7)	0.7928 (7)	0.7932 (7)	0.7929 (7)
8	NextTry		0.7927 (8)	0.7928 (8)	0.7927 (8)	0.7929 (8)
9	ankitbk07		0.7927 (9)	0.7930 (9)	0.7927 (9)	0.7929 (9)
10	Rashfi		0.7868 (10)	0.7870 (10)	0.7868 (10)	0.7869 (10)
11	rohanmainali	Silver	0.7847 (11)	0.7833 (11)	0.7847 (11)	0.7830 (11)
12	prerana3		0.7799 (12)	0.7812 (12)	0.7789 (12)	0.7810 (12)
13	TomalJoy		0.7417 (13)	0.7416 (13)	0.7417 (13)	0.7416 (13)
14	Tanvir_77		0.7405 (14)	0.7414 (14)	0.7406 (14)	0.7411 (15)
15	bidhan_*b*		0.7380 (15)	0.7382 (15)	0.7377 (15)	0.7377 (16)
16	AkshYat		0.7360 (16)	0.7360 (16)	0.7360 (16)	0.7360 (14)
17	akshayyy22		0.7225 (17)	0.7234 (17)	0.7217 (17)	0.7219 (17)
18	ysb	YS	0.6926 (18)	0.6927 (18)	0.6923 (18)	0.6923 (18)
19	Durgeshverma24iitrm	MLP	0.6636 (19)	0.6644 (19)	0.6632 (19)	0.6636 (19)
20	**wangkongqiang**	**wang**	**0.6365 (20)**	**0.6359 (20)**	**0.6360 (20)**	**0.6331 (17)**
21	MDSagorChowdhury	Musafir	0.6179 (21)	0.6862 (19)	0.5828 (21)	0.6233 (18)

10 Limitations of the Work

we are interested in learning about LLMs in computational social science (Thapa et al., 2025b), our paper mainly focuses on making discussions on

Table 9: The Final Leaderboard of Subtask B: Classifying the Targets of Hate Speech.

#	User	Team Name	Recall	Precision	F1	Accuracy
1	wangxiuxian	TUJ-MI	0.6383 (1)	0.6759 (1)	0.6530 (1)	0.6426 (1)
2	Ryuan		0.6204 (2)	0.6556 (2)	0.6335 (2)	0.6426 (1)
3	ray-sushant		0.6021 (5)	0.6169 (4)	0.6057 (3)	0.6305 (2)
4	jiarranDiana	IMU-L	0.6038 (3)	0.6230 (3)	0.6015 (4)	0.6305 (3)
5	Sujal_Maharjan	Multimodal Kathmandu	0.5922 (6)	0.5666 (5)	0.5777 (5)	0.5823 (4)
6	bidhan_cb		0.6032 (4)	0.5407 (10)	0.5628 (6)	0.5703 (5)
7	prerana3		0.5504 (7)	0.5653 (7)	0.5539 (7)	0.5904 (3)
8	ankitbk07		0.5249 (9)	0.6044 (6)	0.5486 (8)	0.5823 (6)
9	shrutigurung	MemeMasters	0.5059 (10)	0.5427 (9)	0.5150 (9)	0.5382 (7)
10	rohanmainali	Silver	0.5422 (8)	0.5092 (12)	0.5018 (10)	0.5181 (8)
11	akshayyy22		0.4869 (11)	0.5289 (11)	0.4984 (11)	0.5151 (10)
12	MDSagorChowdhury	Musafir	0.4143 (12)	0.4008 (13)	0.3739 (12)	0.4418 (10)
13	**wangkongqiang**	**wang**	**0.3322 (13)**	**0.5552 (8)**	**0.3405 (13)**	**0.4779 (9)**
14	Durgeshverma24iitrm	MLP	0.2757 (14)	0.3158 (14)	0.2739 (14)	0.4096 (11)

hate speech for this task. This is because we are quite interested in and good at identifying hate and offense categories in the text (Parihar et al., 2021). Due to our lack of utilization of image features, we are unable to make good use of the image content in the dataset of this sharing task. Also, the model used for extracting text features is similar to that in the sklearn package of CountVectorizer, but it cannot extract the content from the text very well. We believe that by combining a better image feature model, more refined text feature extraction, and conducting appropriate text preprocessing, our model can achieve better results. These are all our future tasks.

11 Ethical Considerations

Our work focuses on hate speech detection and target classification within LGBTQ+ related multimodal content, a domain that is inherently sensitive and requires heightened ethical awareness throughout all research stages. We address the following key ethical concerns:

We used public OCR tools (Google Vision API, pytesseract, EasyOCR) and open-source libraries (e.g., scikit-learn, TensorFlow) to extract and analyze text. We disclose our models, source code, and hyperparameters openly at our GitHub repository, promoting transparency and reproducibility. However, we caution that open-source release of models detecting sensitive content must be accompanied by ethical usage disclaimers and limitations.

We aim to improve our methods by integrating more robust and interpretable models, minimizing biases, and involving domain experts—especially from affected communities—in future annotation and evaluation processes. Ethical AI practices will remain a guiding principle in our ongoing research.

Acknowledgments

We are very grateful to the organizers of the Shared Task on Multimodal Hate, Humor,and Stance Detection in Marginalized Movement@CASE2025 and the School of Information of Yunnan University for providing the environment and equipment.

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.

Rosa Falotico and Piero Quatto. 2015. Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, 49:463–470.

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)*.

Mustafa Oz, Akan Yanik, and Mikail Batu. 2023. Under the shadow of culture and politics: Understanding lgbtq social media activists' perceptions, concerns, and strategies. *Social Media+ Society*, 9(3):20563051231196554.

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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17320–17332, Miami, Florida, USA. Association for Computational Linguistics.

Leslie N. Smith. 2017. Cyclical learning rates for training neural networks. In *Computer Vision and Pattern Recognition*. arxiv.org.

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. 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 Socio-political 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.

Luminaries@CASE 2025: Multimodal Hate Speech, Target, Stance and Humor Detection using ALBERT and Classical Models

Akshay Esackimuthu
Department of Computer Science and Engineering
Sathyabama Institute of Science And Technology
Chennai - 600119 , Tamil Nadu , India
akshayesackimuthu@gmail.com

Abstract

In recent years, the detection of harmful and socially impactful content in multimodal online data has emerged as a critical area of research, driven by the increasing prevalence of text-embedded images and memes on social media platforms. These multimodal artifacts serve as powerful vehicles for expressing solidarity, resistance, humor, and sometimes hate, especially within the context of marginalized socio-political movements. To address these challenges, this shared task introduces a comprehensive, fine-grained classification framework consisting of four subtasks: (A) detection of hate speech, (B) identification of hate speech targets, (C) classification of topical stance toward marginalized movements, and (D) detection of intended humor. By focusing on the nuanced interplay between text and image modalities, this task aims to push the boundaries of automated socio-political event understanding and moderation. Using state-of-the-art deep learning and multimodal modeling approaches, this work seeks to enable a more effective detection of complex online phenomena, thus contributing to safer and more inclusive digital environments.

1 Introduction

Hate speech detection has become an essential component in fostering a safer and more inclusive digital ecosystem. In today's highly connected world, where social media and online platforms shape public discourse, the rapid dissemination of hateful content can lead to severe social and psychological harm, particularly against marginalized communities. Effectively identifying and mitigating such content not only protects vulnerable groups but also promotes constructive dialogue and reduces the risk of conflict escalation.

Recent advancements in natural language processing (NLP) and computer vision (Parihar et al., 2021) have significantly enhanced the capabilities of hate speech detection systems, particularly in multimodal contexts where images are embedded with textual content. By jointly analyzing both modalities, it is possible to capture subtle nuances, such as sarcasm or implied hostility, that would otherwise be missed in unimodal approaches. This is particularly critical in the context of memes and other visual artifacts commonly used to spread hateful or harmful narratives.

In line with this vision, the shared task introduced in CASE 2025 (Thapa et al., 2025) as part of workshop (Hürriyetoğlu et al., 2025) focuses on the detection of hate speech, identification of targeted entities, stance classification towards marginalized movements, and detection of humor in multimodal social media content. Building upon this framework, our study explores the integration of transformer-based models and classical machine learning techniques to tackle these challenges.This analysis has base references from (Thapa et al., 2024) and (Thapa et al., 2023).

Specifically, we employ the ALBERT base transformer model, known for its parameter efficiency and strong performance in semantic understanding tasks. In addition, we incorporate classical models such as XGBoost, LightGBM, Gradient Boosting, and MLP classifiers, which allow for diverse feature perspectives and robust ensembling strategies. Our approach combines traditional feature engineering (e.g., syntactic and TF-IDF features) with deep contextual embeddings to capture both surface-level and deep semantic cues.

Through weighted ensembling and subtask-specific optimizations, we aim to improve the fine-grained detection of hate speech and its associated attributes, ultimately contributing to more effective content moderation and fostering healthier online interactions.

Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025, pages 71–75, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-008

2 Dataset & Task Description

2.1 Overview

In the evolving digital landscape, text-embedded images, such as memes and infographics, have emerged as powerful tools of expression, particularly in social and political discourse. These images often blend textual and visual cues, creating a complex multimodal environment that challenges traditional content moderation and hate speech detection methods. Within the context of the marginalized movement, such images can serve dual roles: amplifying voices of solidarity and simultaneously perpetuating harmful stereotypes or hostility. The nuanced interplay between humor and offense further complicates moderation efforts, as satire often straddles the delicate boundary between critique and hate.

Recognizing this complexity, the shared task CASE2025 proposes a comprehensive classification framework, focusing on four distinct yet interrelated subtasks: detection of hate speech, identification of hate speech targets, classification of stances toward marginalized movement, and humor detection. The data set used for this study consists of meticulously annotated text-embedded images for each subtask, enabling a detailed exploration of online discourse. The dataset is curated from (Shah et al., 2024) and (Bhandari et al., 2023).The features of the dataset is given in the table 1.

Table 1: Features of the dataset

Field	Description
filename	Name of the file with index value
text	Text extracted from text-embedded images
label	Ground truth label or category associated with the text/image

2.1.1 Subtask A: Detection of Hate Speech

The primary objective of this subtask is to determine whether an image contains hateful content. Images are annotated with binary labels: **Hate** and **No Hate**. This binary categorization simplifies initial screening yet serves as a critical foundation for deeper analysis in subsequent subtasks.

Label	Count
No Hate	2,065
Hate	1,985
Total	**4,050**

Table 2: Distribution of labels in Subtask A for binary hate speech detection.

2.1.2 Subtask B: Classification of Targets of Hate Speech

For images identified as hateful, the next step is to pinpoint the specific target of hate. The dataset categorizes targets into four classes: **Undirected**, **Individual**, **Community**, and **Organization**. This fine-grained categorization enables a better understanding of hate speech dynamics and the intended victim groups.

Label	Count
Undirected	617
Individual	199
Community	931
Organization	238
Total	**1,985**

Table 3: Label-wise distribution for Subtask B, focused on hateful images only.

2.1.3 Subtask C: Classification of Topical Stance

This subtask focuses on identifying the stance expressed by the image towards the marginalized movement. Stance classification is crucial for understanding the broader sentiment landscape and distinguishing supportive content from oppositional narratives. The dataset includes three stance labels: **Neutral**, **Support**, and **Oppose**.

Label	Count
Neutral	1,166
Support	1,527
Oppose	1,357
Total	**4,050**

Table 4: Distribution of stances towards the marginalized movement in Subtask C.

2.1.4 Subtask D: Detection of Intended Humor

The final subtask involves determining whether the image is intended to convey humor, sarcasm, or

satire. Humor plays a significant role in shaping public perceptions and often acts as a vehicle for veiled hostility. Detecting such elements is essential for nuanced content moderation. The dataset labels images as **Humor** or **No Humor**.

Label	Count
Humor	2,737
No Humor	1,313
Total	**4,050**

Table 5: Distribution of humor-related labels in Subtask D.

3 Methodologies Used

3.1 Preprocessing

To ensure the textual content extracted from images is clean and analysis-ready, extensive preprocessing steps were implemented:

- Conversion to lowercase to normalize textual patterns.
- Removal of punctuation, stop words, URLs, emojis, and special symbols to minimize noise and irrelevant cues.
- Lemmatization using the NLTK library to reduce words to their base forms, improving semantic understanding.
- Tokenization using built-in mechanisms in TF-IDF and transformer models to prepare the text for vector-based analysis.

3.2 Feature Engineering

Several feature engineering strategies were employed to enhance the representational capacity of the text:

- **TF-IDF vectors** for classical machine learning models, capturing term importance and contextual relevance.
- **Syntactic features**, including:
  - Word count, which helps assess verbosity and potential aggressiveness.
  - Stopword ratio, indicating content density.
  - Frequency of punctuation and uppercase letters, often correlated with emotional intensity.
  - Average word length, providing additional stylistic insights.

3.3 Models Used

Transformer-Based Model: ALBERT The **ALBERT (A Lite BERT) base v2** model was utilized as a primary deep learning approach due to its efficiency and superior performance in text classification tasks. ALBERT leverages self-attention mechanisms to capture complex token relationships, enabling it to understand nuanced semantic and syntactic patterns present in text-embedded images. It was fine-tuned on each subtask-specific labeled dataset, allowing it to adapt to different classification objectives.The flow of the process is shown in Figure 1

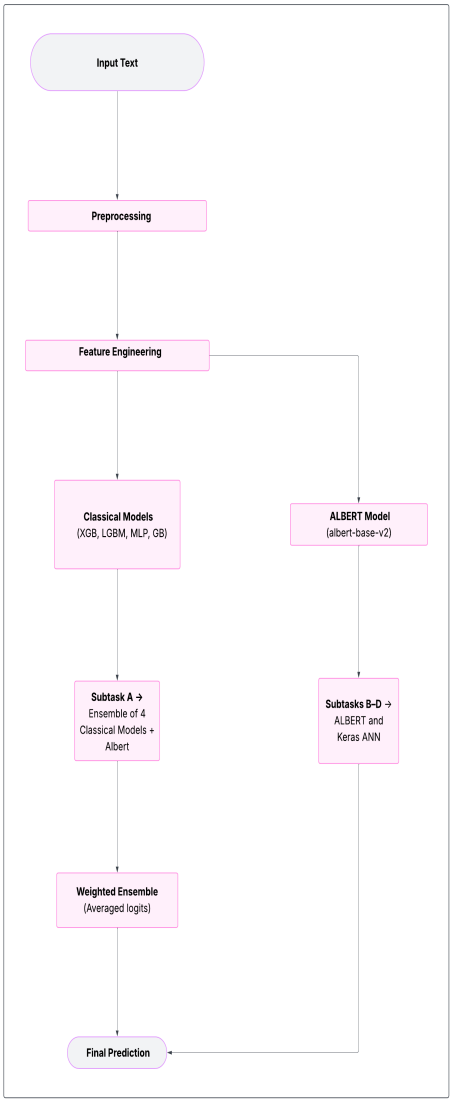

Figure 1: Text-embedded image

4 Results & Discussion

This section presents the implementation details and comprehensive analysis of the results obtained

for each subtask of the CASE2025 Multimodal Hate Speech Detection Shared Task. The evaluation was carried out using standard metrics such as accuracy, precision, recall, and F1-score, and the results are discussed in depth below.

4.1 Subtask A: Hate Speech Detection

For the primary subtask of determining whether a given text contains hate speech, a combination of transformer-based and classical machine learning models was explored. The ALBERT (albert-base-v2) model was fine-tuned using the simpletransformers library, while classical models including XGBoost, LightGBM, GradientBoostingClassifier, and MLPClassifier were trained using TF-IDF and syntactic features. A weighted ensembling approach was adopted to integrate predictions from these models.

The model ensemble achieved an F1-score of **0.7234**, with a recall of **0.7225**, precision of **0.7217**, and accuracy of **0.7219**, securing a competitive rank (14th) on the leaderboard. The results demonstrate that leveraging ensemble strategies can effectively balance the strengths of transformer-based deep representations with classical feature-driven approaches. However, the slight margin for improvement suggests potential benefits from further fine-tuning ensemble weights and incorporating additional linguistic features.

Metric	Score
Recall	0.7225
Precision	0.7217
F1-Score	0.7234
Accuracy	0.7219

Table 6: Subtask A: Hate Speech Detection

4.2 Subtask B: Hate Speech Target Identification

In this subtask, the goal was to classify the target of hate speech into four categories: undirected, individual, community, or organization. The ALBERT model was fine-tuned for multiclass classification, and a separate feedforward ANN was developed using Keras Sequential API.

The ALBERT model achieved an F1-score of **0.4984**, with a recall of **0.4869**, precision of **0.5289**, and accuracy of **0.5542**, ranking 6th. These results highlight the inherent challenge of accurately distinguishing nuanced targets within hate speech. While the transformer model effectively captured contextual dependencies, the relatively lower scores compared to subtask A suggest that future work could incorporate more sophisticated target-specific features or additional multimodal cues.

Metric	Score
Recall	0.4869
Precision	0.5289
F1-Score	0.4984
Accuracy	0.5542

Table 7: Subtask B: Target Identification

4.3 Subtask C: Stance Classification

The task of stance classification involved categorizing posts as hate-supporting, neutral, or counter-hate. The ALBERT model and a Keras-based ANN were trained independently without ensembling.

The ALBERT model yielded an F1-score of **0.5305**, with a recall of **0.5355**, precision of **0.5434**, and an accuracy of **0.5523**, placing 9th overall. These moderate scores indicate the complexity of stance interpretation, which often depends on subtle linguistic cues and contextual nuances. Integrating additional context-aware features or user-level metadata could potentially enhance performance in future iterations.

Metric	Score
Recall	0.5355
Precision	0.5434
F1-Score	0.5305
Accuracy	0.5523

Table 8: Subtask C: Stance Classification

4.4 Subtask D: Humor Detection

In the humor detection subtask, the aim was to determine whether a hateful post contained humorous or sarcastic elements. The ALBERT model and ANN were both trained separately for this binary classification task.

The ALBERT model achieved an F1-score of **0.6070**, recall of **0.6030**, precision of **0.6274**, and accuracy of **0.6844**, resulting in a 15th place ranking. These results underscore the challenge of detecting humor, which is often subjective and culturally dependent. Despite reasonable performance,

further improvement could be obtained by integrating multimodal features such as emoji usage, stylistic patterns, or contextual image data.

Metric	Score
Recall	0.6030
Precision	0.6274
F1-Score	0.6070
Accuracy	0.6844

Table 9: Subtask D: Humor Detection

4.5 Comparative Analysis

Across all subtasks, the ALBERT (albert-base-v2) model consistently outperformed the ANN-based approaches, demonstrating the strong contextual learning capabilities of transformer architectures. While classical models and ANN methods showed promising trends in certain tasks, they generally lagged behind the fine-tuned transformer in overall performance.

The application of preprocessing techniques such as lemmatization, stopword removal, and syntactic feature engineering contributed significantly to model robustness. Furthermore, the ensembling strategy employed in subtask A highlighted the effectiveness of combining diverse models to improve predictive performance.

5 Conclusion

Our approach to the CASE 2025 shared task combined the interpretability of classical machine learning models with the representational power of transformers. Ensembling methods improved performance in hate speech detection (Subtask A), and even single-model approaches worked effectively for the remaining subtasks. Future work includes integrating image features and extending ensemble methods to all subtasks.

Limitations

- We did not incorporate the image modality or multimodal fusion
- Our ensemble approach was limited to Subtask A due to time and resource constraints.
- We did not explore data augmentation or advanced fusion techniques.

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.

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)*.

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)*.

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

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

*These authors contributed equally to this work

Abstract

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

1 Introduction

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

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

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

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

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

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

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

2 Related Work

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

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

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

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

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

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

3 Dataset and Task

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

The shared task is structured into the following four subtasks:

- **Subtask A: Hate Speech Detection** — Identify whether the meme contains hateful content, distinguishing between *hate* and *non-hate* expressions.
- **Subtask B: Target Classification** — Determine the entity or group targeted by the hate speech, categorized as *Undirected*, *Individual*, *Community*, or *Organization*.
- **Subtask C: Stance Classification** — Assess the stance conveyed towards the identified target, classified as *Support*, *Neutral*, or *Oppose*.
- **Subtask D: Humor Detection** — Classify whether the meme is intended to be binary classification of humor, an important dimension given humor's complex role in meme communication.

Analyzing the class distribution reveals mild imbalances, mainly in Subtasks B and Subtasks C, where a large majority of samples are labelled as

neutral or non-targeted. Humor detection subtask, on the other hand, benefits from intentional oversampling and a more balanced representation is reached, enabling fairer model evaluation.

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

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

Table 1: Dataset Sample Distribution Across Subtasks

4 Methodology

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

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

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

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

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

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

4.1 BERTRES Model Architecture

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

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

Table 2: Training Setup for Each Subtask

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

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

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

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

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

4.2 Training Setup

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

5 Results & Discussion

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

Table 3: Final Leaderboard Performance of BERTRES

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

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

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

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

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

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

6 Conclusion

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

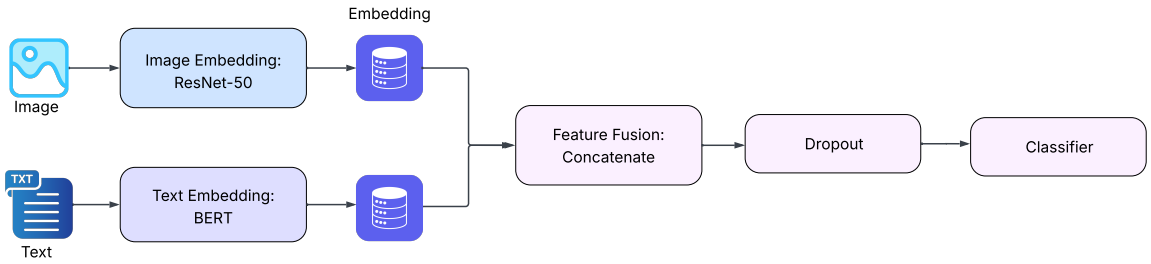

Figure 1: System overview of the BERTRES model combining textual and visual modalities for meme classification across four subtasks.

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

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

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

7 Limitations

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

8 Ethical Considerations

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

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

References

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

Aashish Bhandari, Siddhant B Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. Crisishatemm: Multimodal analysis of directed and undirected hate speech in text-embedded images from russia-ukraine conflict. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1994–2003.

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

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

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

Ali Hürriyetoğlu, Surendrabikram Thapa, and Hristo Tanev. 2025. Findings and insights from the 8th workshop on challenges and applications of automated extraction of socio-political events from text. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of 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 Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

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

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

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

Kritesh Rauniyar, Sweta Poudel, Shuvam Shiwakoti, Surendrabikram Thapa, Junaid Rashid, Jungeun Kim, Muhammad Imran, and Usman Naseem. 2023. Multi-aspect annotation and analysis of nepali tweets on anti-establishment election discourse. *IEEE Access*, 11:143092–143115.

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

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

Shuvam Shiwakoti, Surendrabikram Thapa, Kritesh Rauniyar, Akshyat Shah, Aashish Bhandari, and Usman Naseem. 2024. Analyzing the dynamics of climate change discourse on twitter: A new annotated corpus and multi-aspect classification. In *Joint 30th International Conference on Computational Linguistics and 14th International Conference on Language Resources and Evaluation, LREC-COLING 2024*, pages 984–994. European Language Resources Association (ELRA).

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

Surendrabikram Thapa, Farhan Ahmad Jafri, Ali Hürriyetoğlu, Francielle Vargas, Roy Ka Wei Lee, and Usman Naseem. 2023a. Multimodal hate speech event detection-shared task 4. In *CASE 2023-Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text, associated with 14th International Conference on Recent Advances in Natural Language Processing, RANLP 2023*, pages 151–159. Association for Computational Linguistics.

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

Surendrabikram Thapa, Kritesh Rauniyar, Farhan Ahmad Jafri, Shuvam Shiwakoti, Hariram Veeramani, Raghav Jain, Guneet Singh Kohli, Ali Hürriyetoğlu, and Usman Naseem. 2024b. Stance and hate event detection in tweets related to climate activism-shared task at case 2024. In *7th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text, CASE 2024*, pages 234–247. Association for Computational Linguistics.

Surendrabikram Thapa, Kritesh Rauniyar, Farhan Ahmad Jafri, Hariram Veeramani, Raghav Jain, Sandesh Jain, Francielle Vargas, Ali Hürriyetoğlu, and Usman Naseem. 2024c. Extended multimodal hate speech event detection during russia-ukraine crisis-shared task at case 2024. In *7th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text, CASE 2024*, pages 221–228. Association for Computational Linguistics.

Surendrabikram Thapa, Kritesh Rauniyar, Shuvam Shiwakoti, Sweta Poudel, Usman Naseem, and Mehwish

Nasim. 2023b. Nehate: Large-scale annotated data shedding light on hate speech in nepali local election discourse. In *ECAI 2023*, pages 2346–2353. IOS Press.

Surendrabikram Thapa, Siddhant Bikram Shah, Kritesh Rauniyar, Shuvam Shiwakoti, Surabhi Adhikari, Hariram Veeramani, Kristina T. Johnson, Ali Hürriyetoğlu, Hristo Tanev, and Usman Naseem. 2025. Multimodal hate, humor, and stance event detection in marginalized sociopolitical movements. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2025)*.

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 un-

derstanding. 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 Socio-political 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.

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.

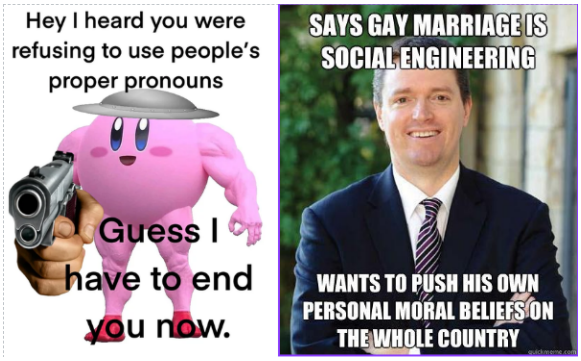

Figure 4: Sub-Task D Training Data Example (Left: No Humor, Right: Humor)

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

Multimodal Deep Learning for Detection of Hate, Humor, and Stance in Social Discourse on Marginalized Communities

Durgesh Verma
IITRAM, Ahmedabad
Gujarat, India
durgesh.verma.24co@iitram.ac.in

Abhinav Kumar
CSED, MNNIT Allahabad
Prayagraj, India
abhik@mnnit.ac.in

Abstract

Internet memes serve as powerful vehicles of expression across platforms like Instagram, Twitter, and WhatsApp. However, they often carry implicit messages such as humor, sarcasm, or offense especially in the context of marginalized communities. Understanding such intent is crucial for effective moderation and content filtering. This paper introduces a deep learning-based multimodal framework developed for the CASE 2025 Shared Task on detecting hate, humor, and stance in memes related to marginalized movements. The study explores three architectures combining textual models (BERT, XLM-RoBERTa) with visual encoders (ViT, CLIP), enhanced through cross-modal attention and Transformer-based fusion. Evaluated on four subtasks, the models effectively classify meme content—such as satire and offense—demonstrating the value of attention-driven multimodal integration in interpreting nuanced social media expressions.

1 Introduction

Memes have emerged as a dominant medium of communication in the digital age, enabling users to express emotions, opinions, and social commentary in humorous yet impactful ways. Their wide dissemination on platforms such as Twitter, Instagram, and WhatsApp makes them not only vehicles of entertainment but also instruments of cultural and ideological expression. Despite their seemingly innocuous appearance, memes can carry coded language, sarcasm, and implicit ideologies that may reinforce hate (Parihar et al., 2021; Roy and Kumar, 2025; Swain et al., 2022; Kumar et al., 2021), misinformation, or discrimination (Zannettou et al., 2018). Their interpretative flexibility often depends on the viewer's cultural background, context, and personal values (Kiela et al., 2020), making automatic intent recognition particularly challenging.

Figure 1: A meme promoting 'Heterosexual Pride Month' — raising concerns about LGBTQ+ exclusion.

What makes memes powerful is also what complicates their analysis: they integrate both visual and textual elements, with meaning frequently emerging from the interaction between the two modalities. A single meme can appear humorous to some while being deeply offensive to others. This inherent ambiguity necessitates sophisticated approaches to computational analysis that can reason across modalities and cultural contexts.

Figure 1 illustrates the importance of multimodal reasoning. This meme, sourced from the CASE 2025 shared task (Thapa et al., 2025a; Hürriyetoğlu et al., 2025), visually depicts a heterosexual family shielding their child from colored rain. While it may appear neutral or protective at first glance, the colored rain can be interpreted as representing LGBTQ+ pride, implying an exclusionary and discriminatory undertone.

Interpretation 1: Neutral Perspective: At first glance, the meme may appear to convey a positive or protective sentiment—parents shielding their child from colorful rain. This can be interpreted as

Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025, pages 98–106, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-012

a metaphor for responsible parenting, without any harmful connotation.

Interpretation 2: Critical Perspective: Upon closer examination, the rainbow-colored rain suggests a symbolic representation of LGBTQ+ identity. The umbrella labeled "Heterosexual Pride Month" implies protection from LGBTQ+ influence, thus reinforcing anti-LGBTQ+ sentiment and promoting a harmful ideological stance.

The latest developments in the impressive deep learning, particularly in the field of multimodal learning, have allowed extraction and reasoning of both textual and visual features. The alignment of vision and language, which is vital in the understanding of the layered semantics of memes, has had an impressive result on the architectures, including CLIP (Radford et al., 2021) and BLIP (Li et al., 2022). Particular to meme analysis, the Hateful Memes Challenge (Thapa et al., 2024) and Memotion Analysis tasks (Sharma et al., 2020) have inspired new studies on multimodal hate (Thapa et al., 2023; Bhandari et al., 2023) and sentiment analysis.

It has been observed that unimodal learning systems cannot capture subtle contextual data; hence, to address the situation, we present a multimodal deep learning framework to integrate visual and textual information in a multi-modal environment by means of the fusion strategy of attention. The given approach can identify not only direct but also expressive forms of hate, such as sarcasm and positions of ideology. Our model follows a similar pattern but uses pre-trained architectures (XLM-R on the text data, CLIP model on the visual information) and colleges a Transformer-based fusion module to enable more robust performance with better interpretability on multiple downstream standardized data sets.

The rest of the sections are organized as follows: Section 2 discusses related work for memes identification, Section 3 discusses proposed methodology in detail, The outcome of the proposed model is listed in Section 4 and Section 5 concludes the paper.

2 Related Work

Detecting harmful or misleading memes presents a significant challenge due to their inherently multimodal structure often requiring nuanced understanding of visual cues and embedded text. Over the years, multiple approaches have been developed to address this challenge, ranging from unimodal to state-of-the-art multimodal architectures.

2.1 Text and Vision Models for Content Moderation

Early approaches primarily focused on either the textual or the visual component of memes. Models like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) were widely used for text analysis, while CNN-based models such as ResNet (He et al., 2016) and DenseNet (Huang et al., 2017) handled image classification. However, these unimodal methods often failed to capture the interaction between image and text, a critical element in meme understanding.

2.2 Multimodal Detection and the Facebook Hateful Memes Challenge

The Facebook Hateful Memes Challenge (Kiela et al., 2020) emphasized the need for multimodal solutions by presenting memes where neither text nor image alone conveyed hate. Transformer-based models such as ViLBERT (Lu et al., 2019) offered early solutions for joint vision-language learning.

The introduction of CLIP (Radford et al., 2021) further advanced this field by aligning visual and textual representations in a shared embedding space. Leveraging this, Hate-CLIPper (Kumar and Nandakumar, 2022) and MemeCLIP (Shah et al., 2024) demonstrated robust performance for hateful meme detection and multi-label classification tasks such as humor, stance, and hate.

2.3 Recent Advances in Meme Understanding

Recent studies have further refined multimodal fusion strategies. Align-before-Attend (Hossain et al., 2024) aligns image and text features prior to fusion to improve hate detection performance, particularly on multilingual datasets. Evolver (Huang et al., 2024) applies prompt-based chain-of-evolution reasoning, enabling the model to use historical meme context for interpreting intent.

LMM-RGCL (Mei et al., 2025) introduces a two-stage contrastive learning approach for fine-tuning large multimodal models and achieves state-of-the-art results across six meme datasets. Lin et al. (Lin et al., 2024) propose an explainable debate framework between LLMs (Thapa et al., 2025b) for modeling conflicting viewpoints within memes. M3Hop-CoT (Kumari et al., 2024) uses a multimodal chain-of-thought strategy to enhance misog-

ynous meme detection performance, especially in datasets like MAMI.

In multilingual settings, Chauhan and Kumar (Chauhan and Kumar, 2025) employ XLM-RoBERTa with ViT and BiLSTM-attention for detecting misogyny in Tamil and Malayalam memes. GuardHarMem (El-amrany et al., 2025) incorporates caption generation with fusion-based detection for improved interpretability and performance (F1 $\approx$ 0.91).

2.4 Research Gap

Despite recent advances in multimodal learning, many existing approaches still rely heavily on task-specific architectures, handcrafted feature engineering, or late fusion strategies that treat textual and visual modalities in isolation until the final stage. These limitations restrict the models' ability to capture fine-grained interactions between modalities and often reduce their generalizability across tasks.

3 Methodology

Each meme sample contains both image and OCR text. All subtasks are multi-class or binary classification problems, requiring both modalities for accurate prediction (see Table 1 for task description). The additional information on the dataset can be found in (Thapa et al., 2025a).

This section presents two multimodal architectures designed for the Shared Task on Multimodal Hate, Humor, and Stance Detection in Marginalized Movements @CASE 2025 (Thapa et al., 2025a; Hürriyetoğlu et al., 2025). Both architectures process meme images and their OCR-extracted text, aiming to predict four semantic properties: hatefulness, targeted group, stance, and humor. The overall task is framed as a multi-task learning problem. The flow diagram for the proposed model can be seen in Figure 2.

Our work introduces a unified transformer-based architecture that leverages early fusion of multimodal features through cross-modal attention. Specifically, we explore three distinct combinations of pre-trained language and vision encoders:

- **XLM-RoBERTa + CLIP**: We concatenate 768-dimensional textual embeddings from XLM-RoBERTa with 512-dimensional text and 512-dimensional image embeddings from CLIP, forming a comprehensive 1792-dimensional multimodal representation.
- **BERT + ViT**: This configuration fuses 768-dimensional text embeddings from BERT (Devlin et al., 2019) with 768-dimensional image embeddings from the Vision Transformer (ViT), resulting in a 1536-dimensional joint feature space.
- **XLM-RoBERTa + ViT**: Here, both text and image features are of 768 dimensions, producing a 1536-dimensional combined representation.

These fused embeddings are then processed through a Multi-Head Attention Transformer (Vaswani et al., 2017) that enables deep interaction between modalities at multiple representation levels. Our models are evaluated across all four subtasks of the CASE 2025 competition to assess their robustness, transferability, and domain-independence. The experimental results validate the effectiveness of our early-fusion attention-based approach in capturing nuanced multimodal intent, outperforming or matching task-specific baselines while maintaining general applicability.

3.1 Architecture 1 : Transformer-based Fusion

This section describes the architecture and training procedure of our unified multimodal classification framework.This approach leverages pretrained encoders for independent modality processing and combines their outputs using a Transformer-based fusion module.

3.1.1 Text Encoding with XLM-RoBERTa

We process the OCR-extracted text from memes using the multilingual transformer XLM-RoBERTa (Conneau et al., 2019). Given an input text sequence $\mathbf{x}_{\text{text}}$, we obtain its contextual representation via the final [CLS] token embedding:

$$\mathbf{h}_{\text{xlmr}} = \text{XLM-R}(\mathbf{x}_{\text{text}})_{\text{[CLS]}} \in \mathbb{R}^{768}$$

This text encoder is fine-tuned using binary cross-entropy loss:

$$\mathcal{L}_{\text{text}} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)$$

where $\hat{y}_i = \sigma(\mathbf{W}_t \cdot \mathbf{h}_{\text{xlmr}} + b_t)$ is the predicted probability of the positive class.

Table 1: CASE 2025 Shared Task Subtasks and Labels

Subtask	Objective	Label Names	Encoded Labels
A	Detect hate speech in the meme.	No Hate, Hate	0, 1
B	Identify the target of hateful memes.	Undirected, Individual, Community, Organization	0, 1, 2, 3
C	Determine stance toward movements.	Neutral, Support, Oppose	0, 1, 2
D	Detect humor/satire/sarcasm.	No Humor, Humor	0, 1

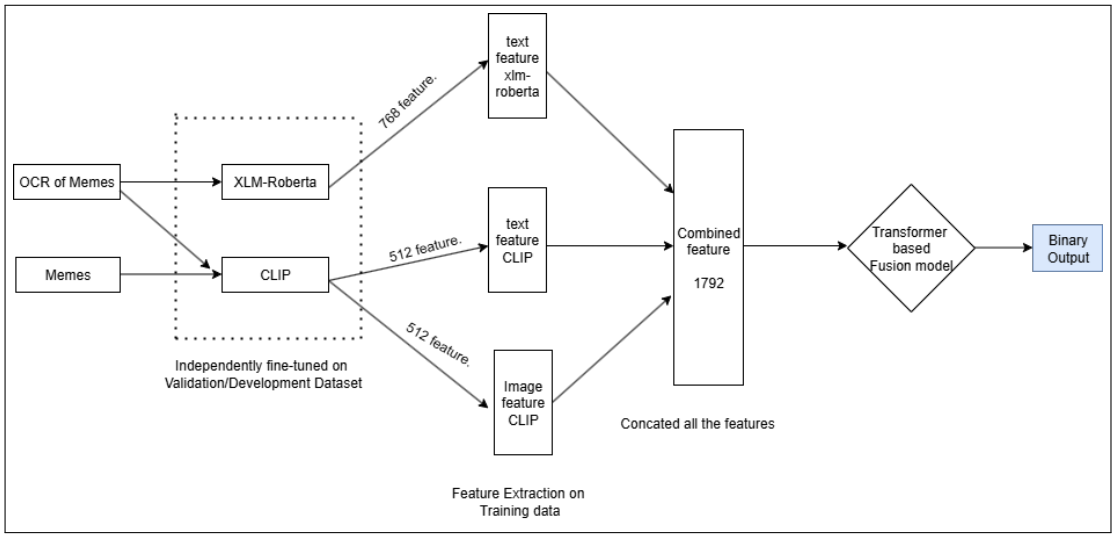

Figure 2: Block diagram of the multimodal architecture for HM, MAMI, and MultiOFF datasets.

3.1.2 Visual and Textual Embedding with CLIP

We employ the Contrastive Language–Image Pretraining (CLIP) model (Radford et al., 2021) to obtain aligned embeddings for both the meme image $\mathbf{x}_{\text{img}}$ and its OCR-extracted text $\mathbf{x}_{\text{ocr}}$. CLIP provides a joint representation:

$$\mathbf{h}_{\text{clip}} = \left[\mathbf{h}_{\text{img}}; \mathbf{h}_{\text{ocr}}\right] \in \mathbb{R}^{1024}, \quad \mathbf{h}_{\text{img}}, \mathbf{h}_{\text{ocr}} \in \mathbb{R}^{512}$$

Both image and text features are extracted via CLIP's pretrained encoders and optionally fine-tuned using:

$$\mathcal{L}_{\text{clip}} = -y \log \sigma(\mathbf{W}_c \cdot \mathbf{h}_{\text{clip}} + b_c)$$

3.1.3 Feature Fusion and Classification

The output representations from XLM-R and CLIP are concatenated into a single vector:

$$\mathbf{z} = [\mathbf{h}_{\text{xlmr}}; \mathbf{h}_{\text{img}}; \mathbf{h}_{\text{ocr}}] \in \mathbb{R}^{1792}$$

This vector is linearly projected to a reduced dimension d for input into a Transformer encoder:

$$\mathbf{z}_{\text{proj}} = \mathbf{W}_p \cdot \mathbf{z} + \mathbf{b}_p, \quad \mathbf{W}_p \in \mathbb{R}^{d \times 1792}$$

The Transformer encoder $\mathcal{T}$ with positional encodings captures inter-modal interactions:

$$\mathbf{z}_{\text{fused}} = \mathcal{T}(\text{PosEnc}(\mathbf{z}_{\text{proj}}))$$

A multilayer perceptron (MLP) followed by a sigmoid activation performs final classification:

$$\hat{y} = \sigma(\text{MLP}(\mathbf{z}_{\text{fused}}))$$

3.1.4 Training Configuration

The fusion classifier is trained using the following setup: Optimizer: Adam, Learning rate: 1×10^{-4}, Loss function: Binary Cross-Entropy, Batch size: 16, and Epochs: 100. The final prediction $\hat{y}$ is thresholded at 0.4:

$$\hat{y} = \begin{cases} 1, & \text{if } \sigma(f(\mathbf{z})) \geq 0.4 \\ 0, & \text{otherwise} \end{cases}$$

3.2 Architecture 2 : Transformer-based Fusion with Bidirectional Cross Attention

3.2.1 Text Encoder: XLM-RoBERTa

The textual content from memes is tokenized and passed into a fine-tuned **XLM-RoBERTa** model. We extract contextual token embeddings $\mathbf{H}_t \in$

$\mathbb{R}^{L\times d}$ and apply mean pooling to obtain the final text embedding:

$$\mathbf{h}_t = \frac{1}{L}\sum_{i=1}^{L}\mathbf{H}_t^{(i)}$$

where L is the sequence length and $d = 768$ is the hidden dimension.

3.2.2 Image Encoder: Vision Transformer (ViT)

The meme image is resized and fed into a pre-trained **ViT** model. The image is split into P patches and encoded to obtain $\mathbf{H}_v \in \mathbb{R}^{P\times d}$. Similar to the text stream, we perform mean pooling:

$$\mathbf{h}_v = \frac{1}{P}\sum_{j=1}^{P}\mathbf{H}_v^{(j)}$$

3.2.3 Bidirectional Cross-Modal Attention

We apply **bidirectional multi-head attention** between visual and textual sequences to model fine-grained interactions:

$$\mathbf{A}_{t\leftarrow v} = \text{MHA}(\mathbf{H}_t, \mathbf{H}_v, \mathbf{H}_v) \quad (1)$$
$$\mathbf{A}_{v\leftarrow t} = \text{MHA}(\mathbf{H}_v, \mathbf{H}_t, \mathbf{H}_t) \quad (2)$$

These attention outputs are pooled to form final fused features:

$$\mathbf{z} = [\text{MeanPool}(\mathbf{A}_{t\leftarrow v}); \text{MeanPool}(\mathbf{A}_{v\leftarrow t})] \in \mathbb{R}^{2d}$$

3.2.4 Multi-task Classification Heads

The fused vector $\mathbf{z}$ is passed through four independent multi-layer perceptrons (MLPs), one for each subtask:

$$\hat{\mathbf{y}}^{(s)} = \text{Softmax}(f^{(s)}(\mathbf{z})), \quad s \in \{A, B, C, D\}$$

Each head uses a cross-entropy loss:

$$\mathcal{L}_{\text{total}} = \sum_s \lambda_s \cdot \mathcal{L}_{\text{CE}}^{(s)}$$

where λ_s are task-specific weights (default 1.0).

3.2.5 Training Configuration

Following hyperparameter were used to train the proposed model: Optimizer: AdamW with weight decay, Learning rate: 3×10^{-5} (with warm-up), Epochs: 30, Loss: Multi-task CrossEntropy, Backbone Freezing: First 5 epochs.

3.3 Architecture 3 : Lightweight Fusion with PCA and Multi-Head Attention

3.3.1 Static Feature Extraction

In this architecture, we use frozen encoders for feature extraction:

- **Text:** BERT [CLS] embeddings $\mathbf{t} \in \mathbb{R}^{768}$
- **Image:** ViT global embeddings $\mathbf{v} \in \mathbb{R}^{768}$

3.3.2 Dimensionality Reduction

We apply PCA separately:

$$\mathbf{t}' = \text{PCA}(\mathbf{t}) \in \mathbb{R}^{128}, \quad \mathbf{v}' = \text{PCA}(\mathbf{v}) \in \mathbb{R}^{128}$$

3.3.3 Multi-Head Attention Fusion

The reduced embeddings are passed through dense layers and fused using multi-head attention:

$$\mathbf{q} = \mathbf{W}_q \cdot \mathbf{v}' \quad \mathbf{k}, \mathbf{v}_{att} = \mathbf{W}_k \cdot \mathbf{t}' \quad (3)$$
$$\mathbf{z}_{\text{fused}} = \text{MHA}(\mathbf{q}, \mathbf{k}, \mathbf{v}_{att}) \quad (4)$$

3.3.4 Multi-task Output and Loss

The attention output is flattened and passed into shared dense layers, followed by task-specific classification heads:

$$\hat{\mathbf{y}}^{(s)} = \text{Softmax}(g^{(s)}(\mathbf{z}_{\text{fused}})), \quad \forall s \in \{A, B, C, D\}$$

We minimize total cross-entropy loss across all subtasks:

$$\mathcal{L}_{\text{total}} = \sum_s \mathcal{L}_{\text{CE}}^{(s)}$$

3.3.5 Training Configuration

Following hyperparameter were used to train the proposed model: Optimizer: Adam, Learning Rate: 1×10^{-4}, Epochs: 50, Loss: Cross-entropy per subtask.

4 Results

Table 2 presents the performance of our multimodal model (XLM-R/BERT + ViT/CLIP) across the four subtasks defined in the CASE 2025 shared task. The model achieves reasonably competitive performance in binary tasks such as Hate Detection (Subtask A) and Humor Detection (Subtask D), obtaining an F1 score of 0.6602 and 0.6564, respectively. These results indicate that the model is

Table 2: Performance of the Best Performed Models for Different Sub-tasks

Tasks	Model	Recall	Precision	F1 Score	Accuracy
Subtask-A	XLM-R + ViT + Attention	0.6614	0.6654	0.6602	0.6627
Subtask-B	BERT + ViT + Attention	0.3158	0.2739	0.4096	0.4217
Subtask-C	XLM-R + ViT + Attention	0.4723	0.4905	0.4674	0.4694
Subtask-D	XLM-R + ViT + Attention	0.6554	0.6575	0.6564	0.7002

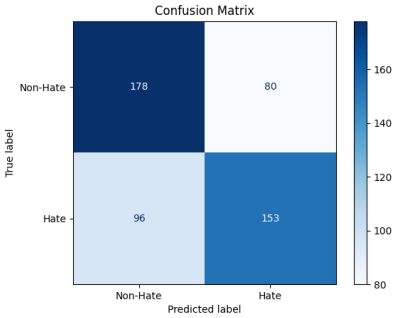

Figure 3: Confusion matrix Subtask A

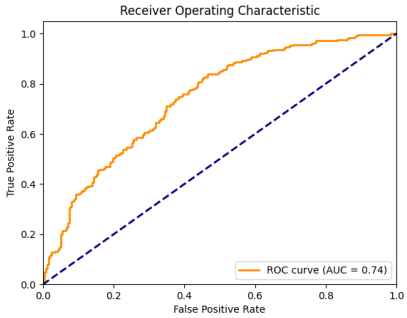

Figure 4: ROC curve multiclass subtask A

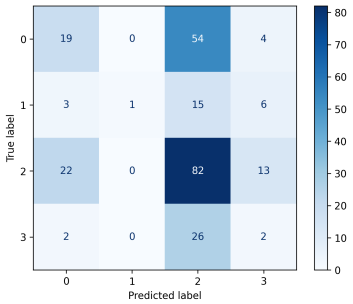

Figure 5: Confusion matrix Subtask B

able to capture surface-level multimodal features to some extent.

However, the performance significantly drops in more semantically complex tasks particularly in Target Identification (Subtask B) and Stance Classification (Subtask C). For instance, the F1 score for Subtask B was only 0.4096, far from the top performer (0.6530). This gap indicates challenges in understanding nuanced and context-specific semantic relationships.

A key reason behind the lower performance in these subtasks is likely due to semantic misalignment between the textual and visual streams. The model often misinterprets the context when the image and text convey conflicting or sarcastic messages. Since memes are frequently designed with contradictory visuals and text (e.g., humorous images paired with hateful text or vice versa), the fusion mechanism occasionally diverges in learning, either overemphasizing the visual cue or misjudging the intended sentiment of the text.

The confusion metric and ROC for for the testing labels for xlm-roberta + VIT+ Attention model can be seen in Figures 3 and 4, respectively. The confusion metric and ROC for for the testing labels for BERT + VIT+ Attention model can be seen in Figures 5 and 6, respectively. The confusion metric and ROC for for the testing labels for clm-roberta + CLIP+ Attention based fusion model can be seen in Figures 7 and 8, respectively. Similarly, confusion matrix and ROC curve for the subtask-D can be seen in Figures 9 and 10, respectively.

These findings echo observations from prior works on multimodal sarcasm and hate detection (Fersini et al., 2022),which highlight that surface-level fusion techniques are insufficient when the modalities encode different or even conflicting semantic signals. Thus, future iterations of the model can benefit from deeper semantic alignment modules or attention-based conflict resolution strategies to better handle such intricacies.

5 Conclusion

The proposed model is validated for four subtasks from the CASE 2025 shared task. The system showed promising performance on binary classification tasks like Hate Detection (F1 = 0.6602) and Humor Detection (F1 = 0.6564), indicating that the model can capture explicit cues from both modalities effectively. However, for more semantically complex subtasks such as Target Group Identification and Stance Classification, the performance was notably lower (F1 = 0.4096 and 0.4674 respectively). These tasks require a deeper understanding of socio-political context and subtle narrative tones, which our current model struggled to generalize. One key limitation identified was the semantic misalignment between image and text. The model often failed to resolve contradictions in multimodal

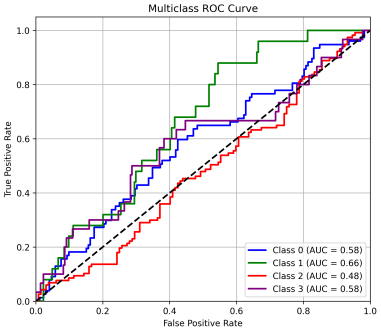

Figure 6: ROC curve for the Subtask B

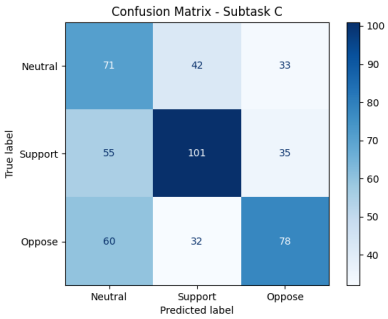

Figure 7: Confusion matrix Subtask C

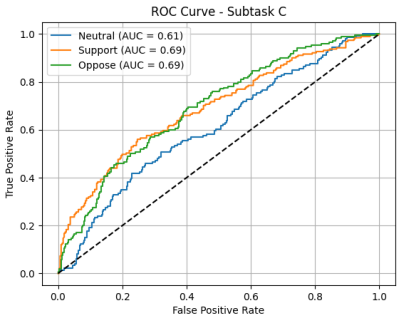

Figure 8: AUC curve multiclass subtask C

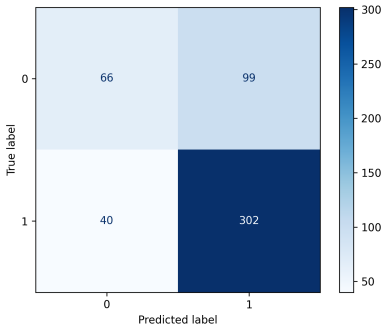

Figure 9: Confusion matrix Subtask D

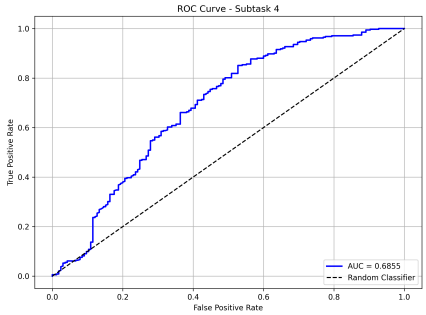

Figure 10: AUC curve multiclass subtask D

memes—where visual irony or sarcasm alters the literal meaning of the text. This led to misinterpretation in scenarios where the intended sentiment was obfuscated through meme-specific humor or design.

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.

Shraddha Chauhan and Abhinav Kumar. 2025. MNLP@DravidianLangTech 2025: transformer-based multimodal framework for misogyny meme detection. In *DravidianLangTech*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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.

Samir El-amrany, Salima Lamsiyah, Matthias R Brust, and Pascal Bouvry. 2025. Guardharmem and harmdetect: a multimodal dataset and benchmlark model for fine-grained harmful meme classification. *Social Network Analysis and Mining*, 15(1):63.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549.

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.

Eftekhar Hossain, Omar Sharif, Moshiul Hoque, and Sarah M. Preum. 2024. Align before attend: Aligning visual and textual features for multimodal hateful content detection. *arXiv preprint arXiv:2402.09738*.

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.

Jinfa Huang, Jinsheng Pan, Zhongwei Wan, Hanjia Lyu, and Jiebo Luo. 2024. Evolver: Chain-of-evolution prompting to boost large multimodal models for hateful meme detection. *arXiv preprint arXiv:2407.21004*.

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.

Abhinav Kumar, Pradeep Kumar Roy, and Sunil Saumya. 2021. An ensemble approach for hate and offensive language identification in english and indo-aryan languages. In *FIRE (Working Notes)*, pages 439–445.

Rohit Kumar and Sathappan Nandakumar. 2022. Hateclipper: Multimodal hateful meme classification using vision-language pretraining. *arXiv preprint arXiv:2210.05916*.

Gitanjali Kumari, Kirtan Jain, and Asif Ekbal. 2024. M3hop-cot: Misogynous meme identification with multimodal multi-hop chain-of-thought. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22105–22138.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM Web Conference 2024*, pages 2359–2370.

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*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Jingbiao Mei, Jinghong Chen, Guangyu Yang, Weizhe Lin, and Bill Byrne. 2025. Robust Adaptation of Large Multimodal Models for Retrieval Augmented Hateful Meme Detection. *arXiv e-prints*, page arXiv:2502.13061.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pam Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning (ICML)*.

Pradeep Kumar Roy and Abhinav Kumar. 2025. Ensuring safety in digital spaces: Detecting code-mixed hate speech in social media posts. *Data & Knowledge Engineering*, 156:102409.

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

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas Pykl, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. Semeval-2020 task 8: Memotion analysis-the visuolingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773.

Manswini Swain, Manish Biswal, Priya Raj, Abhinav Kumar, and Debahuti Mishra. 2022. Hate and offensive language identification from social media: a machine learning approach. In *Electronic Systems and Intelligent Computing: Proceedings of ESIC 2021*, pages 335–342. Springer.

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. 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 Socio-political 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the origins of memes by means of fringe web communities. In *Proceedings of the internet measurement conference 2018*, pages 188–202.

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.

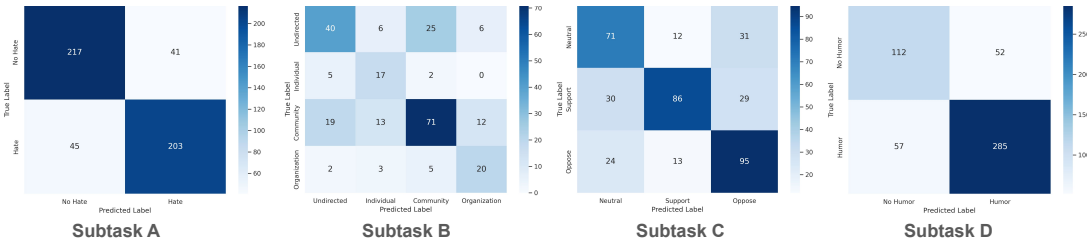

Figure 1: Confusion matrices of Subtasks A, B, C, and D on the evaluation set

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 con-

texts 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)*.

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

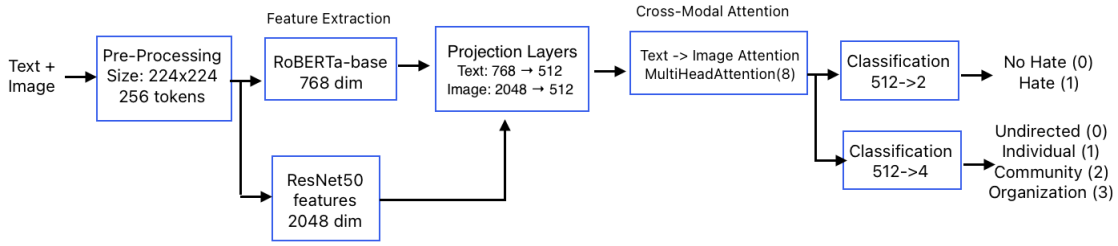

Figure 1: High-level system design for Hate Speech Detection and Target Identification. Text and image inputs are processed through separate encoders (RoBERTa for text and and ResNet50 for images), followed by a cross-modal fusion layer and classification.

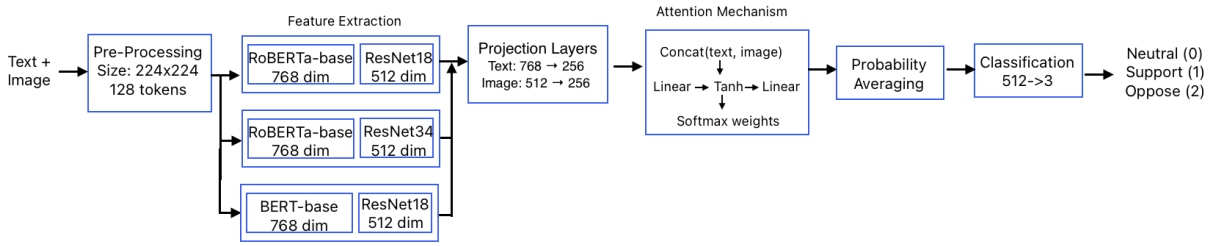

Figure 2: High-level system design for Stance Classification. Three base models (RoBERTa-ResNet18, RoBERTa-ResNet34, and BERT-ResNet18) process multimodal features. Their output probability distributions are averaged to produce the final stance prediction across the three classes: Neutral, Support, and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

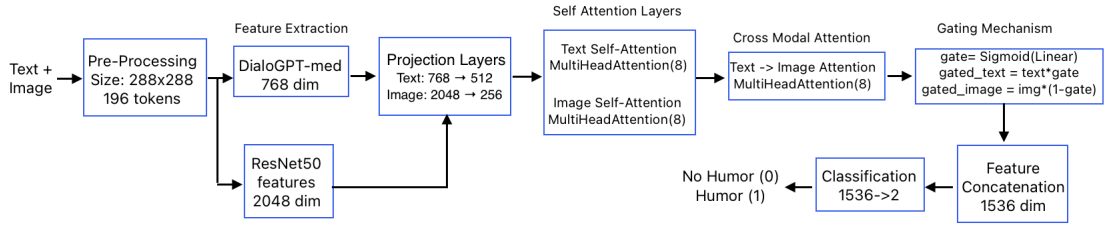

Figure 3: High-level system design for Humor Recognition. DialoGPT-medium for text encoding and ResNet50 for image encoding, followed by cross-modal attention layers fused through a gating mechanism and fed into a multi-layer classifier for binary humor prediction.

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

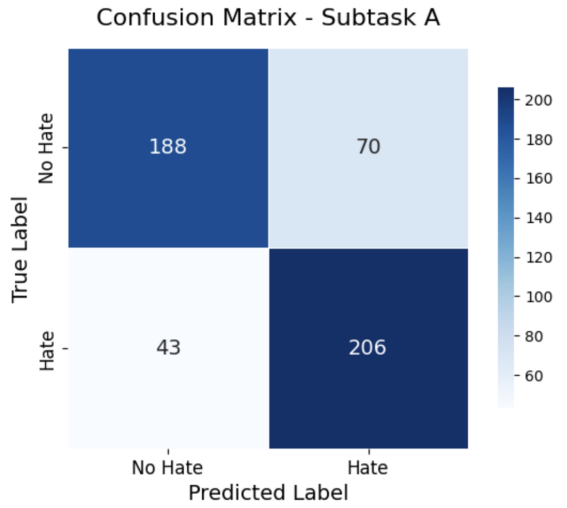

Figure 4: Confusion Matrix for Hate Speech Detection

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.

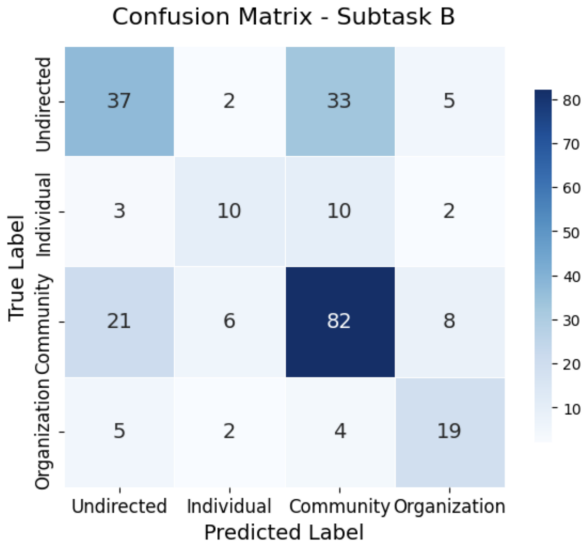

Figure 5: Confusion Matrix for Target Identification

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.

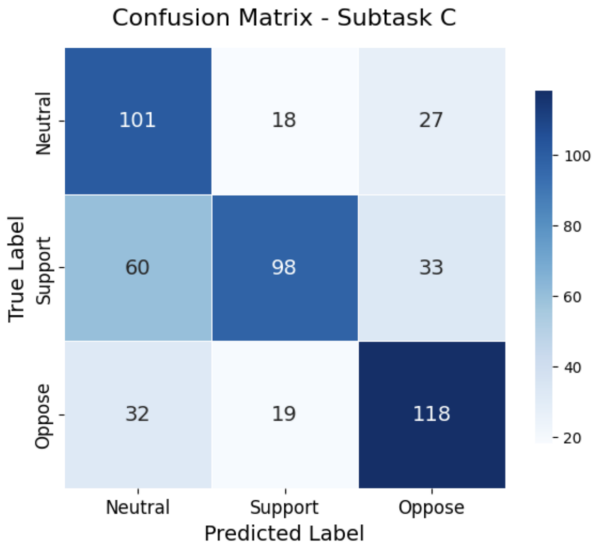

Figure 6: Confusion Matrix for Stance Classification

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.

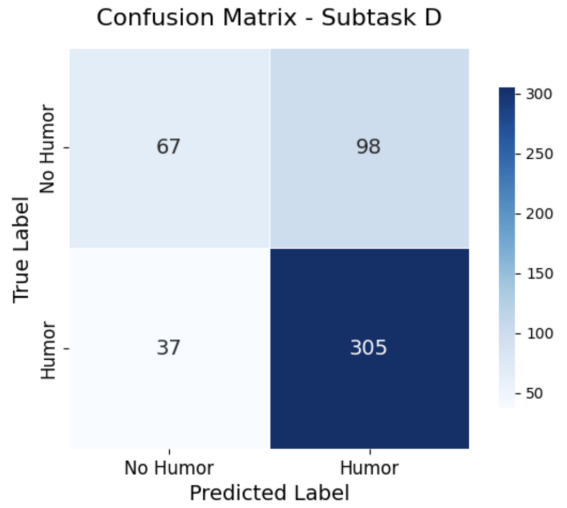

Figure 7: Confusion Matrix for Humor Recognition

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.

TSR@CASE 2025: Low Dimensional Multimodal Fusion Using Multiplicative Fine Tuning Modules

Sushant Kr. Ray[1], **Rafiq Ali**[2], **Abdullah Mohammad**[2], **Ebad Shabbir**[2], **Samar Wazir**[3†]
[1]University of Delhi, [2]Delhi Skill and Entrepreneurship University, [3]De Montfort University
{skray1331, rafworkacc, abdullah90t, ebadshabbir22}@gmail.com,
samar.wazir@dmu.ac.uk

Abstract

This study describes our submission to the CASE 2025 shared task on multimodal hate event detection, which focuses on hate detection, hate target identification, stance determination, and humour detection on text embedded images as classification challenges. Our submission contains entries in all of the subtasks. We propose FIMIF, a lightweight and efficient classification model that leverages frozen CLIP encoders. We utilise a feature interaction module that allows the model to exploit multiplicative interactions between features without any manual engineering. Our results demonstrate that the model achieves comparable or superior performance to larger models, despite having a significantly smaller parameter count. The source code and model checkpoints are available at github.com/sushant-k-ray/FIMIF

1 Introduction

The landscape of digital communication has evolved dramatically with the widespread adoption of social media platforms, fundamentally transforming how individuals express opinions and share content. This evolution has brought significant challenges in content moderation, particularly in the detection of hate speech that increasingly manifests in the form of memes, which are images with text embedded in them used to convey a message. The CASE (Challenges and Applications of Automated Extraction of Socio-political Events from Text) series has consistently addressed these challenges, with recent editions expanding from text-only analysis to encompass multimodal content understanding (Thapa et al., 2023, 2024).

Building upon the success of previous CASE workshops, the multimodal hate event detection task at CASE 2025 (Thapa et al., 2025a; Hürriyetoğlu et al., 2025) represents a natural progression toward addressing more complex multimodal hate speech detection scenarios.

In this paper, we introduce FIMIF (Feature Interaction for Multimodal Integration and Fusion), a model conceptually similar to MemeCLIP (Shah et al., 2024). We utilise modified residual units to leverage the capabilities of deep neural networks while keeping the performance stable. We introduce a feature interaction module that automatically learns exponential and multiplicative relationships between features, enabling the model to capture higher-order interactions. While MemeCLIP is designed for general downstream tasks on meme images, our model specifically targets meme classification. Our approach relies on aggressive compression of multimodal embeddings to very low dimensions, followed by a multiplicative module that allows for richer feature interactions. We provide comprehensive experimental evaluation demonstrating the effectiveness of our approach.

2 Related Works

Hate Speech Detection: The task of hate speech detection has progressed from lexicon-based or shallow machine learning approaches (Burnap and Williams, 2015; Waseem and Hovy, 2016; Davidson et al., 2017) to deep learning models (Parihar et al., 2021). The advent of large pre-trained language models brought significant improvements in hate speech detection. BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DistilBERT (Sanh et al., 2020) introduced contextual embeddings that improved performance on social media hate speech detection. These models have achieved state-of-the-art results on benchmarks such as: HateXplain (Mathew et al., 2021), Offensive Language Identification Dataset (OLID) (Rosenthal et al., 2021), Gab Hate Corpus (Kennedy et al., 2022), and Storm-

†Corresponding author.

Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025, pages 123–132, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-015

front dataset (de Gibert et al., 2018). These language models are very efficient and powerful in terms of language understanding.

Multimodal Tasks: As harmful content increasingly appears in multimodal forms like memes, research has shifted toward models that process both text and images. Datasets such as Facebook's Hateful Memes (Kiela et al., 2020) and MMHS150K (Gomez et al., 2020) have been instrumental in advancing this field. Some recent multimodal hate speech detection datasets include Harm-C (Pramanick et al., 2021a), Harm-P (Pramanick et al., 2021b), DisinfoMeme (Qu et al., 2022), and CrisisHateMM (Bhandari et al., 2023). Early multimodal systems use separate encoders (e.g., ResNet (He et al., 2016) for images and BERT for text) and combine features through concatenation or attention. Later models rely on fusion strategies to combine these different representations.

Vision Language Models: Vision-Language models aim to learn joint representations of visual and textual inputs, typically trained on large-scale image-text pairs. These models are broadly divided into two categories: Dual-encoder models, and Fusion models.

Dual-encoder models, such as OpenAI's CLIP (Contrastive Language-Image Pre-training) (Radford et al., 2021) and Google's ALIGN (Jia et al., 2021), encode images and text separately and align their embeddings using contrastive loss.

CLIP, in particular, has gained popularity due to its strong zero-shot performance and generalisation ability. Trained on 400 million internet image-text pairs, it can embed both modalities into a shared semantic space.

Low-Rank Multimodal Fusion: One of the key challenges in multimodal learning is the integration of information from multiple modalities. While tensor based fusion methods offer powerful and expressive means of capturing interactions between modalities, they are often computationally expensive and suffer from a rapid increase in parameters, particularly when modelling higher-order interactions across multiple input sources (Zadeh et al., 2017).

To mitigate these challenges, Low-rank Multimodal Fusion (LMF) (Liu et al., 2018) has emerged as a scalable and efficient paradigm. Rather than modelling the full tensor representation, LMF approximates it using modality specific low rank projections, which are then combined using element-wise operations. This dramatically reduces the parameter count and computational overhead while still retaining cross-modal interactions. LMF scales linearly with the number of modalities, in contrast to the exponential growth in traditional fusion approaches. We adapt a similar principle with the use of additive and multiplicative layers.

Highway And Residual Networks: Highway networks (Srivastava et al., 2015) and Residual networks (He et al., 2016) are widely used to improve training stability and depth in deep learning models. Residual layers mitigate vanishing gradients by adding skip connections, while highway layers introduce trainable gates to control information passage. These ideas motivate our use of lightweight residual projections to preserve essential features without over-fitting.

Multiplicative Modules: The Neural Arithmetic Logic Unit (NALU) (Trask et al., 2018) introduces a mechanism for learning arithmetic operations in neural networks using log-space computations to model multiplicative relationships. Several variants of NALU have been proposed to improve stability and expressiveness in different settings (Schlör et al., 2020; Madsen and Johansen, 2020; Heim et al., 2020). We extend NALU to multimodal classification in a residual framework to maintain flexibility while modelling higher-order relationships.

3 Dataset And Tasks

Shah et al. (2024) released a novel multimodal dataset, PrideMM consisting of text embedded images for classification of various aspects of hate against marginalised LGBTQ+ movement, and community in online discourse through images, particularly memes. The dataset is divided into four classification tasks: hate detection, hate target identification, stance determination, and humour detection.

The multimodal hate task at CASE 2025 utilises the PrideMM dataset, focusing on discrimination and hate against the LGBTQ+ community. The dataset is divided into an 80/10/10 train-validation-test split. This is different from the PrideMM dataset, where the split is 85/5/10. OCR of the images is provided as supplementary material to aid in the process of classification.

The following table shows the distribution of the training samples:

Task	Label	Samples	%
Hate	No Hate	2065	50.99%
	Hate	1985	49.01%
Target	Undirected	617	31.08%
	Individual	199	10.03%
	Community	931	46.90%
	Organization	238	11.99%
Stance	Neutral	1166	28.79%
	Support	1527	37.70%
	Oppose	1357	33.51%
Humour	No Humour	1313	32.42%
	Humour	2737	67.58%

Table 1: Distribution of the training samples in the shared task dataset.

3.1 Tasks

The PrideMM dataset focuses on following four subtasks:

Subtask A: Hate Detection. This task aims to identify instances of hate speech in the images. This task focuses on identifying whether the images intentionally convey hateful sentiments. The training data is balanced (1.04 : 1), and contains a total of 4050 data samples.

Subtask B: Hate Target Identification. This task focuses on identifying the targets of hate in hateful images. There are four categories: Undirected, Individual, Community, and Organization. Images are labeled 'Undirected' when they target abstract topics, societal themes, or ambiguous targets. Hateful images targeting specific people are labeled 'Individual'. The label 'Community' is used for instances of hate in images targeting broader social, ethnic, or cultural groups. Images targeting corporate entities, institutions, or similar organizations are labeled 'Organization'.

The training data is extremely imbalanced (3.1 : 1 : 4.7 : 1.2), and contains data samples for only those images which convey hate. As a consequence, only 1985 data samples are available for training.

Subtask C: Stance Determination. This task aims to determine the stance that the image is trying to convey towards the topic. There are three categories: Support, Oppose, and Neutral. The 'Support' label is given to images that express support for the goals of the movement, agree with efforts to promote equal rights for LGBTQ+ individuals, or promote awareness of the movement. The 'Oppose' label is given to images that express disagreement with the goals of the movement, deny the problems faced by individuals who identify as LGBTQ+, or dismiss the need for equal rights and acceptance. The 'Neutral' label is given to images that are contextually relevant to the movement but exhibit neither support nor opposition towards the movement.

The training data is fairly well balanced (1 : 1.31 : 1.16), and contains a total of 4050 data samples.

Subtask D: Humour Detection. This task aims to detect whether the image showcases any form of humour, sarcasm, or satire related to the LGBTQ+ pride movement regardless of whether it presents a light-hearted or insensitive perspective on serious subjects.

The training data is imbalanced (1 : 2.08), and contains a total of 4050 data samples.

4 Methodology

In this section, we describe FIMIF (Feature Interaction for Multimodal Integration and Fusion), our proposed model for meme classification. We utilise the CLIP vision-language model to extract multimodal embeddings that effectively encode the semantic content of memes. Figure 1 illustrates the overall architecture of our model. Below, we describe each component in detail.

Pre-Trained CLIP Model: Similar to MemeCLIP, we leverage CLIP encoders for their strong zero-shot generalisation and effective transfer learning capabilities. The CLIP model consists of an image encoder (E_I) and a text encoder (E_T). We freeze the weights of both encoders to retain the knowledge acquired during pre-training. We utilise CLIP ViT-L/14 image encoder pre-trained on 336x336 images instead of 224x224. 336x336 images can better represent high-frequency information than their 224x224 counterparts. Figure 2 presents an example. Note that the use of 336px variant of CLIP's image encoder does not increase the parameter count of the encoder. The unimodal image and text representations $X_I, X_T \in \mathbb{R}^{768}$ effectively encode the semantic content of a meme and are defined as:

$$X_I = E_I(I);\ X_T = E_T(T) \tag{1}$$

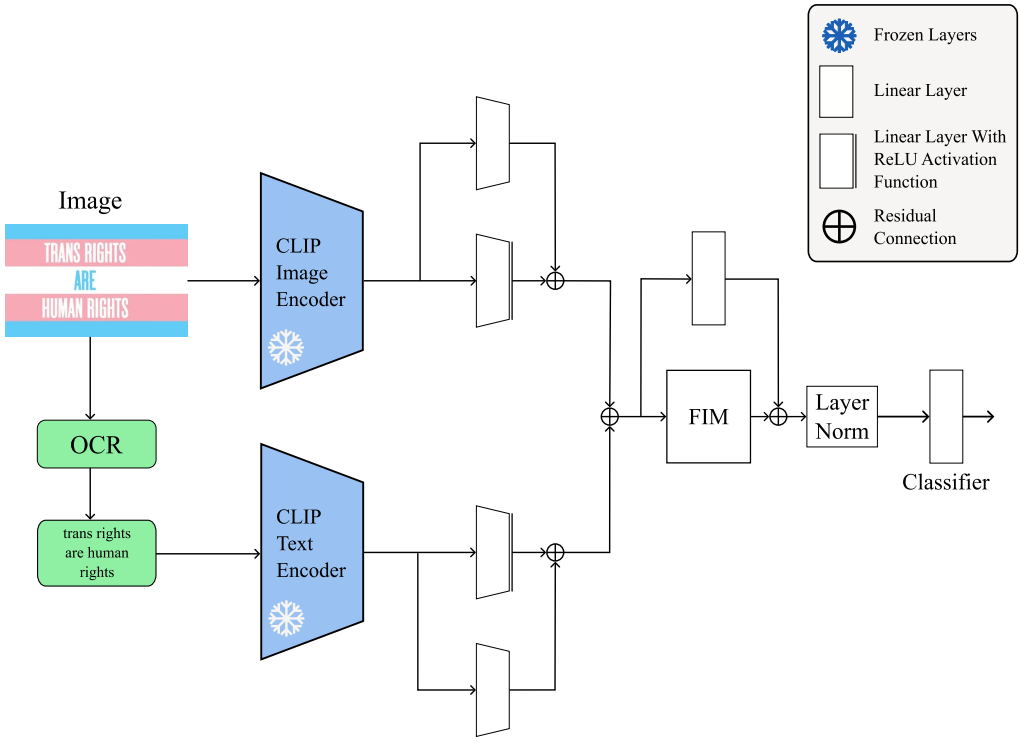

Figure 1: Architecture of our proposed model. Trapeziums are used to represent dimensionality compression.

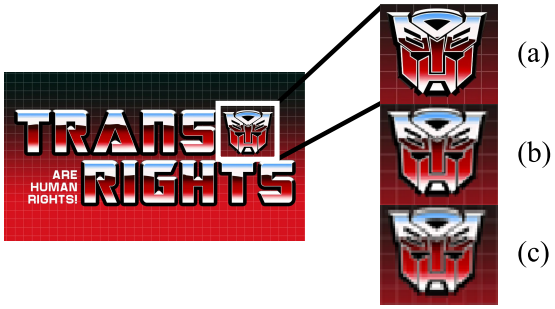

Figure 2: Meme at various resolutions (a) Original resolution (2314x1191) (b) Image downscaled to 336x336 (c) Image downscaled to 224x224. The downscaling and upscaling method used is bicubic interpolation.

where I is the image and T is its corresponding OCR text.

4.1 Linear Residual Projection Layer

Although CLIP is trained to maximise similarity between aligned image-text pairs, the inherently contrastive nature of memes, where visual and linguistic elements often convey conflicting messages, calls for additional adaptation of the embedding space. We hypothesise that only a small subset of elements within the embeddings significantly influence the classification outcome. To capture this, we utilise a modified residual module scheme that effectively compresses the embedding spaces. A regular residual layer is defined as:

$$R(X) = A(X) + X \tag{2}$$

where A is typically a non-linear function. Our modified residual module, although similar to the one described above, performs better in compressing high-dimensional spaces, particularly when combined with lasso regularisation (Tibshirani, 1996). Our residual module is defined as:

$$R(X) = A(X) + B(X) \tag{3}$$

where A is a non-linear function and B is a linear function. The domain and co-domain for both functions are $\mathbb{R}^{768}$ and $\mathbb{R}^{h}$, respectively, where h is a very small number (generally 8, or 16).

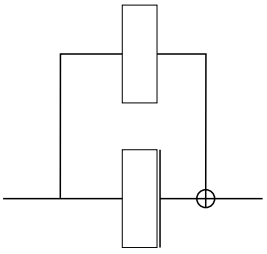

Figure 3: Our Modified Residual Module.

We use a ReLU activation function for A. We utilise this modified residual model for both modalities and combine them to extract linear relations between image and text embeddings. These projection layers result in the bimodal projection $X_{MM} \in \mathbb{R}^h$. Such a layer allows us to leverage the benefits of deep neural network layers while still having the flexibility to use a shallower architecture when required. Our final bimodal residual network is defined as:

$$\begin{aligned} X_{MM} &= R_I(X_I) + R_T(X_T) \\ &= (A_I(X_I) + B_I(X_I)) + \\ &\quad (A_T(X_T) + B_T(X_T)) \end{aligned} \tag{4}$$

4.2 Feature Interaction Module

Since the hidden dimension (h) of the layers is much smaller than CLIP's embedding dimension, we would have difficulty fusing the text and image representations. To capture the non-linear feature interactions in a compact space, we require a multiplicative network. Conceptually, we would like to have a following module:

$$FIM(X) = \begin{bmatrix} M_0(X) \\ M_1(X) \\ \cdots \\ M_{h-1}(X) \end{bmatrix} \tag{5}$$

where,

$$M_i(X) = \prod_{j=0}^{h-1} x_j^{w_{ij}} \tag{6}$$

This module is very generic in nature and can be used for automated feature selection. A module like this, however, would suffer from unstable training due to gradient issues. We design a multiplicative module inspired by Neural Arithmetic Logical Unit (NALU). Rather than directly applying exponentials, we utilise linear arithmetic between inputs in log-space followed by exponentiation. Mathematically, the multiplicative layer can be represented by the following relation:

$$M(X) = B(exp(Wln(ReLU(A(X)) + \epsilon))) \tag{7}$$

where A and B are some linear transformation function with both, domain and co-domain, in $\mathbb{R}^h$. Since logarithm of non-positive numbers is undefined, we use ReLU along with some ϵ (10^{-5} in our case). This strictly positive condition, however, prevents us from multiplying a positive and a negative number. While several variants of the NALU, such as iNALU (Schlör et al., 2020) and NAU (Madsen and Johansen, 2020), introduce complex modifications to address this, we propose a simpler alternative that leverages multiple multiplicative layers. We call this the Feature Interaction Module (FIM). Mathematically, it is defined as follows:

$$FIM(X) = M_a(X) \cdot M_b(X) \tag{8}$$

The Feature Interaction Module is shown in the following diagram:

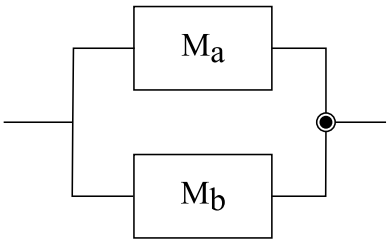

Figure 4: Feature Interaction Module.

We add a residual unit to the FIM in order to allow the network to bypass multiplicative layers if required. The complete residual FIM is defined as:

$$\begin{aligned} F_{MM} &= FIM(X_{MM}) + A(X_{MM}) \\ &= M_a(X_{MM}) \cdot M_b(X_{MM}) + A(X_{MM}) \end{aligned} \tag{9}$$

where A is a linear transformation function with domain and co-domain in $\mathbb{R}^h$.

4.3 Miscellaneous

Classifier: We apply layer normalisation on the outputs of the residual FIM (F_{MM}) before passing it through the classifier. The classifier is a linear transformation function from $\mathbb{R}^h$ to $\mathbb{R}^c$, where c is the number of categories in the given subtask. A softmax function maps the final hidden representations to their respective class probabilities. The predicted class corresponds to the highest probability score.

Class Imbalance: There is a heavy class imbalance in the dataset. To get around this issue, we utilise weighted cross-entropy loss. Further, we utilise minority-class deterministic oversampling for subtask B, where there is an extreme class imbalance. The intuition behind this is to expose the model to more samples from minority classes in order to better classify them. Compared to the high dimensionality of the image and text embeddings

Method	# of trainable Parameters	Hate		Target		Stance		Humour	
		Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
Gradient Boosting	-	78.90	78.90	59.44	57.39	61.54	60.52	76.13	70.60
FIMIF (submission)	25k - 51k	81.85	81.85	63.05	60.57	62.92	62.91	79.68	76.83
FIMIF (best)	25k - 51k	81.85	81.85	64.66	64.61	64.89	64.32	79.68	76.83

Table 2: Classification performance of different models on shared task dataset across two evaluation metrics: Accuracy, and F1 score. The hidden dimension of the FIMIF model for subtask A is set to 16, while a reduced hidden size of 8 is used for all other subtasks.

Method	# of trainable Parameters	Hate			Target			Stance			Humour		
		Acc.	AUC	F1	Acc.	AUC	F1	Acc.	AUC	F1	Acc.	AUC	F1
MemeCLIP	2.6M	76.06	**84.52**	75.09	66.12	**81.66**	58.65	62.00	**80.11**	57.98	80.27	**85.59**	77.21
FIMIF (ours)	25k	**78.11**	83.99	**76.43**	**68.42**	75.97	**62.63**	**63.31**	79.84	**59.52**	**80.47**	85.54	**77.54**

Table 3: Classification performance of different models on PrideMM dataset across three evaluation metrics: Accuracy, AUC, and F1 score. Performance metrics for MemeCLIP is sourced from its corresponding paper. The best performance is highlighted in **bold**.

from the CLIP encoders (768 dimensions each), the size of the training set for subtask B is relatively small, consisting of only 1985 samples. This type of high-dimensional data struggles to generalise. Algorithm 1 presents the pseudocode used for upsampling the minority class.

Algorithm 1 Deterministic Class-wise Upsampling.

Require: Dataset D of (x, y) pairs, number of classes C

Initialise `class_samples`$[0 \dots C-1] \leftarrow$ empty lists
for all $(x, y) \in D$ **do**
  Append (x, y) to `class_samples`$[y]$
end for
$M \leftarrow \max_{c \in \{0,\dots,C-1\}}$ length of `class_samples`$[c]$
`upsampled_dataset` $\leftarrow$ empty list
for $c = 0$ to $C-1$ **do**
  `samples` $\leftarrow$ `class_samples`$[c]$
  $n \leftarrow$ length of `samples`
  $r \leftarrow \lfloor M/n \rfloor$
  for $i = 1$ to r **do**
    Append all elements of `samples` to `upsampled_dataset`
  end for
end for
Shuffle `upsampled_dataset`
return `upsampled_dataset`

Weight Initialisation: All weights and biases in our model are initialised using the Kaiming-Uniform distribution (He et al., 2015), which helps maintain a stable gradient flow during the initial phases of training. However, for the weight matrix W in eq. 7, we instead use the identity matrix as the initial weights. By using this initialisation, we ensure that the multiplicative interactions introduced by the FIM initially behaves in a linear and interpretable manner. This allows the model to gradually learn multiplicative behaviour only when it is beneficial, rather than being forced into a multiplicative domain from the beginning. This identity initialisation improves the performance and training time by converging in the early training stages. We use the identity and zero matrices as layer-norm initial weights and biases, respectively.

5 Results

We provide results of our model on the test set of the respective subtask in table 2. We use weighted gradient boosting as a baseline for its excellent generalisation capability with high dimensional data.

Method	Acc.	AUC	F1
CLIP	**81.62**	**88.87**	**79.89**
BERT	80.43	88.08	78.90
RoBERTa	77.27	87.95	75.38
DeBERTaV3	79.45	87.94	77.71

Table 4: Classification performance on subtask A (hate) validation set with our model on CLIP's ViT-L/14@336px image encoder and different text encoders.

Along with our submission results, we have also provided the best results we have encountered so far in order to demonstrate the viability of these very low parameter models. Table 3 compares our results on PrideMM dataset against MemeCLIP

Method	Hidden Dim.	Hate			Target			Stance			Humour		
		Acc.	AUC	F1	Acc.	AUC	F1	Acc.	AUC	F1	Acc.	AUC	F1
CLIP ViT-L/14 (Image Only 224x224)	8	77.87	86.70	75.82	38.71	64.77	36.27	61.46	79.80	59.64	78.85	84.94	71.87
	16	79.84	87.40	78.40	55.65	68.67	48.66	58.50	79.33	55.96	**80.04**	84.78	74.82
CLIP ViT-L/14@336px (Image Only 336x336)	8	79.05	87.62	77.12	39.92	65.51	37.41	63.44	81.06	60.20	78.46	**85.54**	74.05
	16	81.03	87.90	79.39	56.85	69.31	49.16	61.46	81.14	59.60	79.64	85.28	**75.85**
CLIP ViT-L/14@336px + OCR Text	8	**83.20**	88.72	**81.75**	**60.48**	**70.77**	**51.06**	**64.03**	**81.65**	**62.10**	79.45	85.02	74.75
	16	81.62	**88.87**	79.89	47.58	66.43	43.27	**64.03**	81.20	61.68	79.64	84.45	74.65

Table 5: Our experiments with the use of CLIP's image encoders on the validation set of shared task dataset. We use three evaluation metrics: Accuracy, AUC, and F1 score. The best performance is highlighted in **bold**.

Method	Hidden Dim.	Hate			Target			Stance			Humour		
		Acc.	AUC	F1	Acc.	AUC	F1	Acc.	AUC	F1	Acc.	AUC	F1
FIMIF	8	83.20	88.72	81.75	60.48	70.77	51.06	64.03	81.65	62.10	73.72	84.86	70.69
	16	81.62	88.87	79.89	47.58	66.43	43.27	64.03	81.20	61.68	81.23	85.03	75.77
- FIM	8	79.05	89.18	76.98	59.68	70.76	50.28	62.65	81.73	60.45	76.68	85.22	73.34
	16	81.23	89.51	79.59	59.27	70.43	51.52	63.44	81.54	60.89	76.68	85.01	73.03
- Upsampling	8	79.05	89.18	76.98	58.47	71.16	50.84	63.44	81.33	60.91	75.10	85.07	71.17
	16	81.23	89.51	79.59	60.48	71.44	50.21	61.46	81.24	58.72	77.87	85.12	73.08
- Weighted Loss	8	78.66	89.04	76.54	60.48	72.26	51.87	61.66	81.22	57.75	80.63	85.22	73.92
	16	82.02	89.45	80.25	58.87	71.64	49.10	60.08	81.34	57.26	79.64	85.47	74.46

Table 6: Ablation experiments performed on the validation set of given shared task dataset.

Hidden Dim.	Acc.	AUC	F1
4	81.23	88.75	79.24
8	**83.20**	88.72	**81.75**
16	81.62	88.87	79.89
32	77.27	88.86	75.41
64	81.03	**89.22**	79.10
128	80.24	88.31	78.64
256	68.77	89.11	63.14

Table 7: Classification performance on the subtask A (hate) validation set across different hidden dimensions of our model. The best performance is highlighted in **bold**.

Method	Acc.	F1
MOMENTA (Pramanick et al., 2021b)	83.82	82.80
PromptHate (Cao et al., 2022)	84.47	-
Pro-Cap (Cao et al., 2023)	85.03	-
MemeCLIP (Shah et al., 2024)	84.72	83.74
FIMIF (ours)	**87.01**	**83.94**

Table 8: Performance comparison of meme classification models on the HarMeme-C dataset (binary classification). The best performance is highlighted in **bold**.

on all four subtasks. We compare CLIP's text encoders with other large language models in table 4. These models are trained in a deterministic manner (having no randomness) in order to compare different methods. CLIP's text encoder, despite having a shorter context length of 77 tokens, performs better than BERT, RoBERTa, and DeBERTaV3 (He et al., 2023), each supporting a context length of up to 512 tokens. Table 5 compares the results of our model on CLIP ViT-L/14 224px and 336px image encoders on the validation set of the shared task dataset. Table 7 presents a comparison of our model across different hidden dimensions, showing little to no improvement as the dimension size increases, possibly due to over-fitting. Table 8 reports results on the HarMeme-C dataset (Pramanick et al., 2021a), where our model is compared against several state-of-the-art approaches.

5.1 Ablation Study

We have performed our ablation study on the validation set. We compare our model with the one where feature interaction module has been replaced with a linear transformation layer having a non-linear ReLU activation function. The findings in table 6 suggest that the CLIP embeddings of PrideMM dataset is very linear in nature. Due to its residual design, our implementation of feature interaction module is very generic. It can perform just as well, if not better, than a residual module even when the data does not exhibit multiplicative relationships. The difference between these architectures is likely due to the overhead incurred by having a larger number of parameters (3.5 times that of a residual module). Use of upsampling does not seem to have a significant improvement in performance.

Our upsampling scheme should not have any effect on subtasks A and C, where the worst class ratio is less than 2:1. Any difference is likely due to a different shuffling than their non-upsampling counterparts. The use of weighted loss seems to degrade the performance in tasks B and D. However, the difference is not significant.

6 Conclusion

We present FIMIF (Feature Interaction for Multimodal Integration and Fusion), a lightweight parameter-efficient model that leverages CLIP encoders for multimodal meme classification on PrideMM dataset. Our approach relies on aggressive dimensionality compression. A key finding from our ablation study is that the classification problem becomes mostly linear in nature after this compression, indicating that the dimensionality reduction itself is a critical component of our model's success. Our work highlights the potential of low-dimensional fusion as a viable path toward creating more efficient and sustainable models for complex multimodal tasks.

7 Limitations

Dependence On OCR Quality: The textual input relies heavily on the quality of the OCR. Errors in OCR, such as misread words or missing characters, are directly passed to the text encoder without correction or filtering. Moreover, CLIP's text encoder has a maximum context length of 77 tokens. This severely limits our model's ability to classify text-heavy memes. However, Table 5 indicates that the model achieves comparable performance even without OCR.

Lack Of Future Proofing: The world of memes on the Internet evolves rapidly. Words, images, and cultural references can shift in meaning over time. Since our model heavily relies on the frozen CLIP embeddings, it severely limits the ability of our model to adapt to emerging slangs, visual styles, and evolving socio-cultural contexts. This static representation may cause the model's performance to degrade over time.

8 Ethical Considerations

Environmental Impact: Training deep learning models can have a significant environmental impact, mainly due to high energy consumption and the resulting carbon emissions. To address this, we designed our model with a very low parameter count, which helps reduce the overall computational load. In practice, the most time-consuming step is the extraction of CLIP embeddings, while the actual training phase is relatively quick and lightweight. Our fine-tuning approach helps the model adapt quickly to new datasets, reducing the need for repeated or prolonged training.

Potential For Misuse: Any technology designed to understand and identify a specific type of content can potentially be used for malicious purposes. A model that learns the constituent elements of hateful memes could be used to generate new, more effective hateful content to systematically find loopholes in other detection systems.

Societal Impact Of Automated Moderation: The integration of automated moderation systems into digital platforms introduces several ethical concerns with severe societal implications. While such systems enable scalable and timely identification of harmful content, they also risk amplifying existing biases and disproportionately impacting certain user groups (Thapa et al., 2025b). Models trained on imbalanced or culturally narrow datasets may inadvertently silence marginalised communities, misclassify context-dependent expressions, or fail to generalise across linguistic and cultural boundaries. Automated moderation often lacks transparency and interpretability, limiting users' ability to understand or contest moderation decisions. This opacity can undermine fairness and accountability, particularly in high-stakes environments where content removal may affect public discourse or individual reputation.

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.

Pete Burnap and Matthew L. Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242.

Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. 2023. Procap: Leveraging a frozen vision-language model for hateful meme detection. pages 5244–5252.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2022. Prompting for multimodal hateful meme classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 321–332, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11.

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.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1459–1467.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, page 1026–1034, USA. IEEE Computer Society.

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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

Niklas Heim, Tomas Pevny, and Vasek Smidl. 2020. Neural power units. In *Advances in Neural Information Processing Systems*, volume 33, pages 6573–6583. Curran Associates, Inc.

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)*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Brendan Kennedy, Mohammad Atari, Aida M Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Gwenyth Portillo-Wightman, Shreya Havaldar, Elaine Gonzalez, and et al. 2022. The gab hate corpus.

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. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

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.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256, Melbourne, Australia. Association for Computational Linguistics.

Andreas Madsen and Alexander Rosenberg Johansen. 2020. Neural arithmetic units. In *International Conference on Learning Representations*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

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, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021a. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796, Online. Association for Computational Linguistics.

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

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.

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. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2021. SOLID: A large-scale semi-supervised dataset for offensive language identification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 915–928, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Daniel Schlör, Markus Ring, and Andreas Hotho. 2020. inalu: Improved neural arithmetic logic unit. *Frontiers in Artificial Intelligence*, Volume 3 - 2020.

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

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Highway networks.

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. 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 Socio-political 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.

Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Andrew Trask, Felix Hill, Scott E Reed, Jack Rae, Chris Dyer, and Phil Blunsom. 2018. Neural arithmetic logic units. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

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, San Diego, California. Association for Computational Linguistics.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark. Association for Computational Linguistics.

PhantomTroupe@CASE 2025: Multimodal Hate Speech Detection in Text-Embedded Memes using Instruction-Tuned LLMs

Farhan Amin, Muhammad Abu Horaira, Md. Tanvir Ahammed Shawon,
Md Ayon Mia, Muhammad Ibrahim Khan
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u2004068, u2004029, u1904077,u1804128}@student.cuet.ac.bd, muhammad_ikhan@cuet.ac.bd

Abstract

Memes and other text-embedded images are powerful tools for expressing opinions and identities, especially within marginalized socio-political movements. Detecting hate speech in this type of multimodal content is challenging because of the subtle ways text and visuals interact. In this paper, we describe our approach for Subtask A of the Shared Task on Multimodal Hate Detection in Marginalized Movement@CASE 2025, which focuses on classifying memes as either Hate or No Hate. We tested both unimodal and multimodal setups, using models like DistilBERT, HateBERT, Vision Transformer, and Swin Transformer. Our best system is the large multimodal model Qwen2.5-VL-7B-Instruct-bnb-4bit, fine-tuned with 4-bit quantization and instruction prompts. While we also tried late fusion with multiple transformers, Qwen performed better at capturing text-image interactions in memes. This LLM-based approach reached the highest F1-score of 0.8086 on the test set, ranking our team 5th overall in the task. These results show the value of late fusion and instruction-tuned LLMs for tackling complex hate speech in socio-political memes.

1 Introduction

Social media has become a fast-paced platform where content spreads instantly, with memes playing a big role in communication. But they are also used to spread harmful messages, including hate speech targeting marginalized groups. This kind of content can make online spaces unsafe.Since it is impossible to manually keep up with everything being shared, an automated system has become essential for managing such content. This paper addresses Subtask A of the Shared Task on Multimodal Hate Detection in Marginalized Movement@CASE2025, focusing on binary classification ('No Hate' vs. 'Hate') across a dataset of 4,675 text-embedded images. We draw inspiration from (Parihar et al., 2021), which explores natural language processing for identifying harmful content, shaping our approach to this challenge. To address this challenge, we developed a practical approach by fine-tuning transformer models such as DistilBERT and HateBERT to capture textual nuances, and enhancing it with late fusion to integrate visual data, where Qwen2.5-VL-7B-Instruct-bnb-4bit showed strong capability in interpreting the socio-political nuances of memes. This aligns with (Thapa et al., 2025b), which highlights LLM potential in social science, and builds on (Thapa et al., 2023) and (Chhabra and Vishwakarma, 2024) for multimodal insights. Through this work, we hope to contribute towards more scalable and fair content moderation solutions. Our main contributions are as follows:

- A systematic comparison of unimodal, late-fusion multimodal, and LLM-based architectures for meme hate speech detection.
- An efficient fine-tuning strategy that combines LoRA with 4-bit quantization to adapt a large multimodal LLM under resource constraints.
- Empirical analysis of model predictions, illustrated with representative examples drawn from different regions of the confusion matrix.

2 Related Works

Previous research on multimodal hate speech detection has explored many creative ways to tackle the challenges of online conversations, especially in complex social and political contexts. Early work like (Parihar et al., 2021) used natural language processing to spot hate speech by looking at language patterns that show harmful intent. Later studies, such as (Kashif et al., 2023), used ensemble learning to combine features from different data types for better results. Similarly, (Sahin et al., 2023) improved text analysis by adding syntactic and entity-

Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025, pages 133–138, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-016

level information with transformer models. In another approach, (Aziz et al., 2023) proposed a hierarchical fusion method with separate transformer encoders, and (Chhabra and Vishwakarma, 2024) developed a scalable multilevel attention framework that has influenced our work. While these studies built a strong base for cross-modal hate detection, many still require heavy computation and can be hard to interpret, especially when handling satire or cultural references in memes.

Shared tasks have also helped shape this field by providing benchmarks and valuable datasets. The Shared Task on Multimodal Hate Detection in Marginalized Movement@CASE2025, described in (Thapa et al., 2025a), focuses on binary hate classification, with insights from (Hürriyetoğlu et al., 2025) showing how it has grown. The CASE 2024 shared task (Thapa et al., 2024) featured participants demonstrating the utilization of transformer models like BERT, RoBERTa, and XLNet, as well as effective approaches such as vision transformers and CLIP, which contributed to the outstanding outcomes in hate event detection. The CASE 2023 shared task (Thapa et al., 2023) laid the groundwork for multimodal hate speech detection by combining textual and visual features in text-embedded images

Our dataset comes mainly from (Shah et al., 2024) and its CLIP-based representations, supported by (Bhandari et al., 2023)'s CrisisHateMM work, which highlights the value of careful data curation. Finally, (Thapa et al., 2025b) discusses how large language models can help in social science research, encouraging us to tackle ongoing challenges like telling satire apart from hate in fast-changing socio-political memes.

3 Task and Dataset Description

We have utilized the dataset provided for Subtask A of the Shared Task on Multimodal Hate Detection in Marginalized Movement@CASE2025, as outlined in (Thapa et al., 2025a), focusing on detecting hate speech in text-embedded images. The dataset is divided into training, validation, and test sets with 3,662, 506, and 507 samples, respectively, primarily comprising memes and similar online images. Each image is labeled with a binary tag: 'Hate' or 'No Hate', as detailed in Table 1. This dataset, curated for the 2025 task, serves as our primary resource, with its development informed by (Shah et al., 2024) for CLIP-based representations and supplemented by (Bhandari et al., 2023) for CrisisHateMM analysis, which also shapes the annotation schema.

Table 1: Distribution of images for Subtask A meme hate speech detection.

Dataset	No Hate	Hate	Total
Train	1930	1732	3662
Val	258	248	506
Test	258	249	507

The relatively small dataset size presents challenges for training transformer models, as it increases the risk of overfitting, which motivated our use of data augmentation.

4 Methodology

4.1 Preprocessing

As this is a multimodal task, we have preprocessed both text and image. For the text, we have removed URLs, HTML tags, emojis, and extra whitespace to reduce noise, and converted all text to lowercase for consistency. On the image side, all samples were converted to RGB, resized to 224×224 pixels, and normalized using ImageNet mean and standard deviation to match the input requirements of pretrained models. In total, we have found that 3,662 out of 4,050 training samples had images that matched the text, and we have discarded the rest. All 506 validation samples and all 507 test samples had no missing images.

4.2 Augmentation

To improve model generalization and reduce overfitting, we have applied data augmentation techniques during training. Each image was randomly flipped horizontally and cropped with padding to introduce variation while preserving semantic content. These augmentations were applied only to the training set, while the validation and test sets were left unchanged to ensure consistent evaluation.

4.3 Transformer-based Approach

4.3.1 Unimodal Approach

For the unimodal text classification task, we fine-tuned two transformer-based models: DistilBERT-base-uncased and GroNLP/HateBERT. We selected these models for their pretrained knowledge of general language and hate speech domains. Text sequences were tokenized with a maximum length of 128 tokens. We included a dropout rate of 0.2 in the hidden and attention layers to help prevent overfitting. We trained the models using the Adam

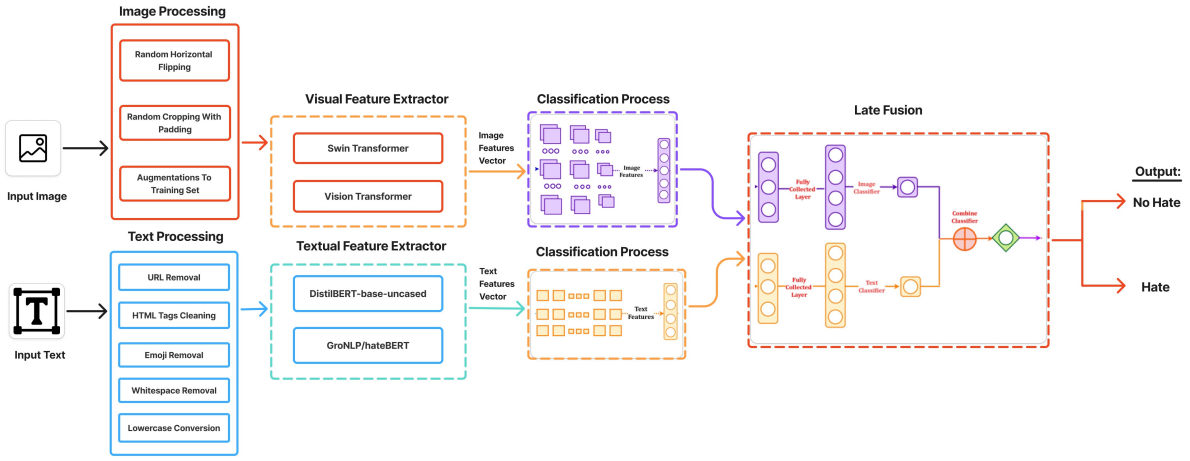

Figure 1: Transformer-based multimodal late fusion architecture for meme hate speech detection

optimizer with a learning rate of 2×10^{-5}, a weight decay of 0.01, and ran the training for 10 epochs on the provided training set. We tokenized the validation and test sets in the same way for evaluation.

For the image-only models, we experimented with Vision Transformer (ViT-base-patch16-224) and Swin Transformer (Swin-T-patch4-window7-224), both of which were first trained on ImageNet. We resized the input images to 224×224, normalized them with standard ImageNet statistics, and converted them to tensors. Each model extracted a 768-dimensional feature vector from the images. We added a dropout layer and a classification head to predict binary labels for Hate and No Hate. Both models were trained for 10 epochs using the AdamW optimizer with a learning rate of 2×10^{-5} and a batch size of 16.

4.3.2 Multimodal Approach

Building on the unimodal baselines, we developed two multimodal architectures that combine both text and images: Swin Transformer with DistilBERT, and HateBERT with Vision Transformer (ViT). In both configurations, 768-dimensional embeddings were extracted separately from the image and text inputs. We combined the embeddings using a late-fusion approach by simply concatenating them. This was followed by a dropout layer to reduce overfitting and a final linear layer for classification. Among the two, the Swin + DistilBERT combination consistently achieved the best performance on the test set.

Late fusion performed better as it enabled the model to process images and text independently before combining their representations, allowing each modality to contribute its strengths. This separation allowed each type of data to focus on its strengths, like visual features from the image and contextual meaning from the text. Combining them later helps the model pick up on subtle clues that come from both. This is really important in hate speech detection, where sometimes the meaning hides in the image, sometimes in the text, and often in both together.

4.4 LLM-Based Approach

We employed a multimodal large language model, `Qwen2.5-VL-7B-Instruct-bnb-4bit`, fine-tuned using the Unsloth framework with 4-bit quantization to improve training efficiency. The goal was to detect hate speech in memes by analyzing both their visual and textual content together. Each training instance was structured as a chat-style conversation, where the user provides an instruction along with a meme image, and the assistant outputs either `0` or `1`, indicating the absence or presence of hate speech, respectively.

We fine-tuned the model over 7 epochs with a batch size of 32. We used LoRA-based fine-tuning (Low-Rank Adaptation) with a rank of 128 applied to both vision and language components. During inference, we applied a zero-shot prompting strategy by employing the same instruction without any meme-specific customization and constrained the model to generate a single classification token.

Our approach achieved a test F1-score of 0.8086, demonstrating efficient performance without relying on handcrafted prompts. This highlights how effective and scalable instruction-tuned multimodal large language models are for detecting hate speech.

Input:
Role: user
Content:
type: text
text: Analyze this meme image and its text content to determine if it contains hate speech.
Respond with '0' if the content is not hateful, or '1' if it contains hate speech.
type: image
image: <PIL.Image meme_image>

Output:
Role: assistant
Content:
type: text
text: '0' or '1'

Figure 2: Prompt provided to Qwen2.5-VL-7B-Instruct-bnb-4bit for meme hate speech detection.

We selected Qwen2.5-VL because it is open-source with efficient fine-tuning support, provides strong multimodal reasoning comparable to larger closed-source models such as GPT-4V, and supports quantized training, where we adopted 4-bit due to hardware limitations with 8-bit.

5 Results and Analysis

5.1 Comparative Analysis

Between the two unimodal text classifiers, HateBERT outperformed DistilBERT-base-uncased, achieving a higher F1-score of 0.7810 compared to 0.7424. This indicates that HateBERT is more effective for hate speech detection on meme texts, likely because it is pretrained specifically on hate speech data. For the unimodal image models, Swin Transformer outperformed ViT, achieving 0.6668 compared to 0.6166, indicating stronger visual feature extraction.

In the multimodal setups, we adopted a late fusion strategy to combine textual and visual representations. Using this approach, Swin Transformer + DistilBERT achieved an F1-score of 0.7790, slightly outperforming the ViT + HateBERT model which scored 0.7576. These results highlight how late fusion enables each modality to contribute its strengths independently before combining them for final prediction, leading to better performance than unimodal baselines.

Finally, the best overall performance was obtained by fine-tuning Qwen2.5-VL-7B-Instruct-bnb-4bit using the Unsloth framework. This model

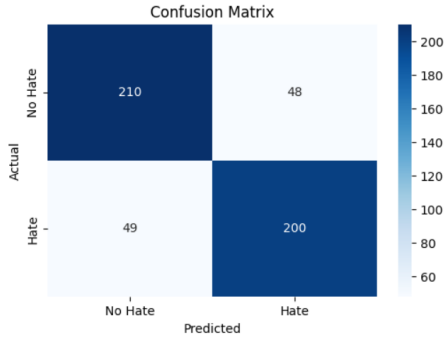

Figure 3: Confusion matrix showing the proposed model's binary classification performance for meme hate speech detection.

achieved an F1-score, precision, and recall of 0.8086. These results highlight the strong potential of large-scale instruction-tuned multimodal models in capturing subtle and cross-modal patterns in hateful memes, outperforming both traditional and multimodal baselines. The results are detailed in Table 2.

5.2 Error Analysis

To better understand the limitations of our best model, we examined its confusion matrix. The fine-tuned Qwen2.5-VL-7B-Instruct-bnb-4bit model correctly predicted most instances in both classes, but some misclassifications remain. It falsely labeled 48 non-hateful memes as hateful and misclassified 49 actual hateful ones.

These errors suggest that while the model performs well overall, it occasionally struggles with subtle or ambiguous cases where hateful intent is not explicit.

5.3 Quantitative Analysis

The confusion matrix shows a fairly balanced distribution of errors, with 48 false positives and 49 false negatives. This indicates that the model is not heavily biased toward one class. However, the nearly equal misclassifications suggest that the model may still be relying on surface-level features, such as specific keywords or visual patterns, rather than understanding the deeper context. Exploring techniques like attention visualization or feature attribution could help reveal what the model is focusing on and guide improvements in handling more nuanced or borderline cases.

5.4 Qualitative Analysis

To further investigate the model's decision patterns, we sampled representative examples from each con-

Table 2: Performance comparison of different models for meme hate speech detection.

Classifier	P	R	F1	Accuracy
Unimodal (Text)				
HateBERT	0.7810	0.7811	0.7810	0.7811
DistilBERT-base-uncased	0.7424	0.7424	0.7424	0.7424
Unimodal (Image)				
Vision Transformer (ViT)	0.6166	0.6116	0.6085	0.6134
Swin Transformer	0.6668	0.6661	0.6660	0.6667
Multimodal				
Swin Transformer + DistilBERT (Late Fusion)	0.7790	0.7768	0.7792	0.7792
ViT + HateBERT (Late Fusion)	0.7576	0.7576	0.7579	0.7579
LLMs				
Qwen2.5-VL-7B-Instruct-bnb-4bit	**0.8086**	**0.8086**	**0.8086**	**0.8087**

fusion matrix category:

Table 3: Example predictions illustrating each category of the confusion matrix.

Category	Example (Image)
True Positive (TP)	1040.png
True Negative (TN)	1155.png
False Positive (FP)	1011.png
False Negative (FN)	1002.png

As shown in Table 3, the model performs well when hateful intent is **clear and explicit**. For instance, it correctly labels a public gathering with signs promoting fairness and unity as **No Hate**, and it also identifies **explicit hostility** in text-based images, such as content expressing negativity toward a music genre.

The model struggles more with memes that are **ambiguous or context-dependent**. A false positive example, a meme satirizing corporate behavior during awareness campaigns, was incorrectly flagged as **Hate**, showing difficulty in separating satire from genuine hostility. Similarly, a false negative case, a humorous meme about dating, was misclassified as **No Hate**, reflecting the challenge of detecting humor that may conceal harmful undertones.

Overall, these patterns highlight the need for stronger cross-modal reasoning and better interpretability to handle subtle and context-driven cases.

6 Conclusion

In this study, we tackled the challenge of detecting hate speech in text-embedded images as part of the Shared Task on Multimodal Hate Detection in Marginalized Movement@CASE2025, focusing on Subtask A. We used a fine-tuned Qwen2.5-VL-7B-Instruct-bnb-4bit model combined with a late-fusion strategy to merge textual and visual features. This approach achieved a solid F1-score of 0.8086 on the test set, highlighting the model's ability to capture subtle interactions between modalities for spotting hate speech in complex memes and socio-political contexts. When compared with unimodal and other multimodal baselines, our method showed clear improvements, especially when humor and harmful messages are mixed together. Overall, our findings provide a useful approach for handling multimodal content, particularly where it relates to marginalized groups.

7 Limitations

We selected Qwen2.5-VL with 4-bit quantization because it is openly available, resource-efficient, and feasible within our computational constraints. However, stronger models (e.g., BLIP-2, LLaVA, GPT-4V) and higher-precision training could potentially yield better results. A single fixed prompt was used for simplicity, though alternative prompting strategies or retrieval-based methods may improve robustness. We adopted late fusion for efficiency, but more advanced cross-modal fusion techniques could capture interactions more effectively. Finally, the dataset (4,675 samples) is rela-

tively small, which may limit generalization and reduce coverage of subtle or context-dependent hate speech, highlighting the need for larger datasets in future work.

8 Ethics Statement

We have been committed to ethical practices in developing a system to detect hate speech in images related to marginalized movements. We understand the risks of mislabeling content and worked to balance false positives and negatives, achieving a strong F1-score. Using a public dataset without extra annotation, we respected privacy and data guidelines. Our goal is to promote safer online spaces by reducing harmful content, while recognizing that human oversight is needed to handle context and avoid bias.

References

Abdul Aziz, Md. Akram Hossain, and Abu Nowshed Chy. 2023. Csecu-dsg@multimodal hate speech event detection 2023: Transformer-based multimodal hierarchical fusion model for multimodal hate speech detection. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2023)*, pages 101–107, Varna, Bulgaria. INCOMA Ltd.

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.

Anusha Chhabra and Dinesh Kumar Vishwakarma. 2024. Mhs-stma: Multimodal hate speech detection via scalable transformer-based multilevel attention framework. *arXiv preprint arXiv:2409.05136*.

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

Mohammad Kashif, Mohammad Zohair, and Saquib Ali. 2023. Lexical squad@multimodal hate speech event detection 2023: Multimodal hate speech detection using fused ensemble approach. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2023)*, pages 84–91, Varna, Bulgaria. INCOMA Ltd.

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.

Umitcan Sahin, Izzet Emre Kucukkaya, Oguzhan Ozcelik, and Cagri Toraman. 2023. Arc-nlp at multimodal hate speech event detection 2023: Multimodal methods boosted by ensemble learning, syntactical and entity features. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2023)*, pages 71–78, Varna, Bulgaria. INCOMA Ltd.

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. 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 Socio-political 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.

ID4Fusion@CASE 2025: A Multimodal Approach to Hate Speech Detection in Text-Embedded Memes Using ensemble Transformer based approach

Tabassum Basher Rashfi, Md. Tanvir Ahammed Shawon, Md. Ayon Mia, Muhammad Ibrahim Khan
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
rashfi2004@gmail.com, {u1904077, u1804128}@student.cuet.ac.bd, muhammad_ikhan@cuet.ac.bd

Abstract

Identification of hate speech in images with text is a complicated task in the scope of online content moderation, especially when such talk penetrates into the spheres of humor and critical societal topics. This paper deals with Subtask A of the Shared Task on Multimodal Hate, Humor, and Stance Detection in Marginalized Movement@CASE2025. This task is binary classification over whether or not hate speech exists in image contents, and it advances as Hate versus No Hate. To meet this goal, we present a new multimodal architecture that blends the textual and visual features to reach effective classification. In the textual aspect, we have fine-tuned two state-of-the-art transformer models, which are RoBERTa and HateBERT, to extract linguistic clues of hate speech. The image encoder contains both the EfficientNet-B7 and a Vision Transformer (ViT) model, which were found to work well in retrieving image-related details. The predictions made by each modality are then merged through an ensemble mechanism, with the last estimate being a weighted average of the text- and image-based scores. The resulting model produces a desirable F1-score metric of 0.7868, which is ranked 10 among the total number of systems, thus becoming a clear indicator of the success of multimodal combination in addressing the complex issue of self-identifying the hate speech in text-embedded images.

1 Introduction

The emergence of online platforms and social media has changed the channels of communication and sharing of ideas basically. At the same time, this unprecedented liberty of speech has triggered a worrying rise in online hate speech—a message through which a person or group of people are verbalized and violated because of their identity (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018). Hate speech, especially that carried in the form of images embedded in text (memes), presents a significant challenge for content moderation and online discourse (Gomez et al., 2020). This combination of the textual and the visual mode of presentation makes detection extremely difficult because in many cases when both textual and visual contents are taken together, they can greatly alter their meaning. With the growing complexity of the phenomenon, there has been a trend of automated hate speech detection in research, which has identified both possible applications and limitations in this area, particularly in Natural Language Processing (NLP)-based methods (Parihar et al., 2021).

It is even more difficult to detect hate speech when humor, satire, or coded language are used in memes to mask hateful intentions. The combination of cultural background and rapidly adapting trends online makes it even more difficult and necessitates the usage of multimodal

Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025, pages 139–145, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-017

explanations that encompass both explicit and implicit indications. The Shared Task on Multimodal Hate, Humor, and Stance Detection in Marginalized Movement@CASE25, in particular, subtask A, which aims to detect the existence of hate speech in images embedded in texts, is a relevant site of discussing these issues (Thapa et al., 2023). Recent work shows that large language models (LLMs) are reshaping computational social science and discourse analysis, while posing key methodological and ethical challenges (Thapa et al., 2025b).

This work presents our approach to Subtask A, where we combine state-of-the-art models from both Natural Language Processing (NLP) and Computer Vision (CV) to create a multimodal system for hate speech detection. The framework has three elements:

1. The textual features are extracted using RoBERTa and HateBERT transformers and classified.

2. EfficientNet and Vision Transformer (ViT) models are used to study the visual content of an image, allowing the identification of visually harmful or offensive content.

3. A scheme of ensemble learning is used to combine the predictions made by any individual modality and thus makes use of complementary information across domains and enhances accurate overall predictions.

The following GitHub repository contains the complete implementation details: https://github.com/RashfiTabassum/Multimodal-Hate-Speech-Detection/tree/main.

2 Related Works

Research on the detection of hate speech in multimodal settings has gone down several approaches with their own limitations. (Pamungkas et al., 2020) achieved 75-80% precision in misogyny detection through the use of machine learning but without visual data. (Derbentsev et al., 2022) concentrated only on text-based methods, whereas (Fortuna and Nunes, 2018) acknowledged that there are few powerful multimodal techniques. (Rawat et al., 2024) explored recent trends, but their techniques struggled with diverse linguistic and visual contexts, reducing generalization. (Kiela et al., 2021) reached an F1 score of 0.80 with the Hateful Memes dataset, which dealt with problems of contextual heterogeneity and uneven distributions. (Cuervo and Parde, 2022) Cuervo and Parde used CLIP to do standardization but had a problem of OCR noise and low flexibility. (Jahan and Oussalah, 2023) restricted their systematic review to NLP-only detection. Meanwhile, (Aluru et al., 2025) introduced a deep-learning framework, yet dependence on the unbalanced information and non-described fusion methods limited its universality.

The CASE shared works have contributed a lot in this field. CASE 2023 (Thapa et al., 2023) was focused on the Russia-Ukraine crisis through the CrisisHateMM dataset, with new subtasks related to hate speech identification and target identifications with multimodal fusion. The scope of CASE 2024 (Thapa et al., 2024) was extended to radicalism, adopting transformer-based NLP and vision models like CLIP and ViT with fusion mechanisms to take into account context, bias, and covert hate such as humor and sarcasm. These two shared tasks provided the foundation for our study.

3 Task and Dataset Description

The Shared Task on Multimodal Hate, Humor, and Stance Detection in Marginalized Movement@CASE2025 (Thapa et al., 2025a; Hürriyetoğlu et al., 2025) has three differ-

ent datasets in Subtask A: Detection of Hate Speech. It contains 3,662 images, including 1,732 hate images and 1,930 no-hate images. The validation entails 506 images, including 248 labelled as Hate and 258 labelled as No Hate. The test set consists of 507 images, and 249 of them are labeled as Hate, whereas 258 belong to No Hate.

Table 1: Distribution of data for Hate and No Hate categories.

Sets	Hate	No Hate	Total
Train	1985	2065	3662
Val	248	258	506
Test	249	258	507

The dataset is based on the Memeclip corpus (Shah et al., 2024) and the CrisisHateMM (Bhandari et al., 2023) dataset, whose annotation schema was modified for this task. These are the core of the CASE 2025 dataset curation.

4 Methodology

The task objective is to determine the occurrences of hate speech in images embedded with text; thus, a multimodal deep learning methodology to be able to utilize the interaction between the visual and linguistic domains is required. To do this, our method uses a multimodal deep learning architecture that combines CNN-based models for images and pretrained transformer models for text, then employs a fusion strategy that capitalizes on the advantages of both modalities.

4.1 Preprocessing

The preprocessing of the textual data is done by removing the URLs, mentions, non-ASCII characters, digits, and excessive white spaces; all tokens will be automatically transformed to lowercase. Comments that are empty are substituted with an already defined placeholder. Pictures will be resized to 224 x 224 pixels and augmented (additional attempts) by rotations, flipping horizontally, color jittering, and cropping of random parts of images. The images are center-cropped, and the statistical parameters of ImageNet are used to normalize them in a consistent way during validation and testing.

4.2 Text-Based Modeling

For the text modality, we fine-tune two pretrained transformer models, RoBERTa-base and HateBERT (GroNLP), to classify text as either Hate or No Hate. They then tokenized the input text via their respective RoBERTaTokenizer and HateBERTTokenizer with their total length truncated to a maximum of 256 tokens. The AdamW optimizer was used with the learning rate of 1×10^{-5}, and training was done in seven epochs. To handle class imbalance, the class weights were calculated using the scikit-learn function compute_class_weight. To get the final probability of prediction of the text modality, the results of both fine-tuned models were averaged:

$$\text{TextProb} = 0.5 \times \text{RoBERTa} + 0.5 \times \text{HateBERT}.$$

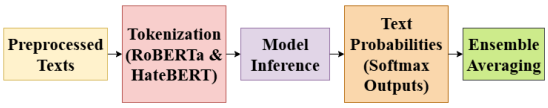

Figure 1: Unimodal Architecture for Text Classification using RoBERTa and HateBERT, followed by Ensemble Averaging.

4.3 Image-Based Modeling

In the case of the image modality, we applied two convolution-based and transformer-based networks: EfficientNet-B7 and Vision Transformer (ViT-B/16). Both of them were pretrained on ImageNet and then fine-tuned on the target dataset, where data augmentation, i.e., horizontal flipping, rotation, color jittering, and random cropping, was used to enable

them to generalize better on unseen data. The training was carried out in 7 epochs using the Adam optimizer at the rate of 1×10^{-5} and 1×10^{-4} of ViT and EfficientNet, respectively. After convergence, the models produced probabilities at the class level; the individual ones were averaged to arrive at the final image prediction:

$$\text{ImageProb} = 0.5 \times \text{EfficientNet} + 0.5 \times \text{ViT}.$$

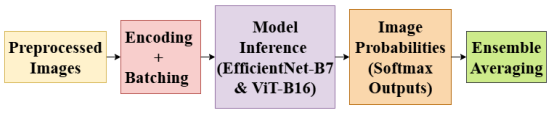

Figure 2: Unimodal Architecture for Image Classification using EfficientNet-B7 and ViT-B16, followed by Ensemble Averaging.

4.4 Multimodal Fusion

We use a late-fusion architecture considering complementary textual and visual data. Textual modalities provide, on average, stronger cues to hate speech as analyzed through the validation procedure, which is empirical. In order to balance the two modalities, we used experiments that changed weights of text-image elements, viz. (0.5, 0.5), (0.7, 0.3), (0.8, 0.2) and (0.9, 0.1). These weight configurations were systematically tested on the validation set in terms of Accuracy, Macro F1, ROC-AUC and class-wise F1 scores. The weighting scheme with 0.7 and 0.3 respectively to textual and visual modality returned the highest Macro F1 and was thus used as the final weighting. The resulting fusion is given as:

$$\text{FinalProb} = 0.7 \times \text{TextProb} + 0.3 \times \text{ImageProb}.$$

Predictions are made by applying a 0.5 threshold on the final probability, classifying the image as Hate or No Hate.

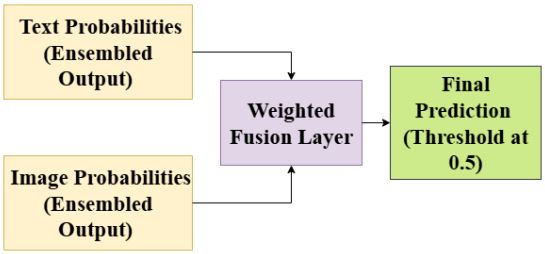

Figure 3: Fusion Layer for Multimodal Prediction using Text and Image Probabilities with Final Thresholding.

5 Experiments and Results

The comparative outputs of various models in terms of macro-averaged Precision (Pr), Recall (Re), and F1-score (F1) have been provided in Table 2. RoBERTa became the best among the text-based models with a macro F1-score of 0.7505, exceeding the results of HateBERT 0.7494. Additional improvement of precision to 0.7990 was made by the ensemble model (RoBERTa + HateBERT), which shows the high ability to combine both models to achieve greater performance. In the image-based models, ViT performed better compared to EfficientNet-B7 with a Macro F1-Score of 0.6351, which is higher than 0.5757 obtained by EfficientNet-B7. The combination of EfficientNet-B7 & ViT had a Macro F1-Score of 0.6311, which shows that two models are more advantageous. The Multimodal Fusion Model, which unites RoBERTa, HateBERT, EfficientNet-B7, and ViT using weights of 70 percent text and 30 percent image, was able to surpass all past models by a big margin. Among all the classification models, the Fusion Model generated the max value of Macro F1-Score (0.7868), Precision (0.7870), and Recall (0.7868).

Table 2: Performance Comparison of Unimodal and Multimodal Models on the Test Dataset

Classifier	Precision	Recall	Macro F1-Score
Unimodal (Text)			
RoBERTa	0.7709	0.7028	0.7505
HateBERT	0.7460	0.7430	0.7494
Ensemble (RoBERTa + HateBERT)	0.7990	0.6546	0.7466
Unimodal (Image)			
EfficientNet-B7	0.5691	0.5622	0.5757
ViT (Vision Transformer)	0.6212	0.6586	0.6351
Ensemble (EfficientNet-B7 + ViT)	0.6220	0.6345	0.6311
Multimodal (Late Fusion)			
Fusion of RoBERTa, HateBERT, EfficientNet-B7, and ViT (70% Text, 30% Image)	0.7870	0.7868	0.7868

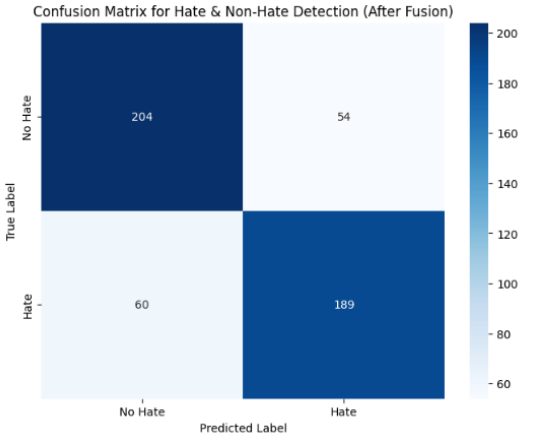

Figure 4: Confusion Matrix for Hate and Non-Hate Detection after Multimodal Fusion

6 Error Analysis

Figure 4, a confusion matrix, indicates some essential misclassification patterns and gives many insights concerning the behavior of the model and its limitations. The multimodal fusion model shows strong results (204 total true negatives and 189 true positives), but there is a tendency to misclassify "No Hate" content as "Hate" (54 false positives) and "Hate" content as "No Hate" (60 false negatives). Such mistakes indicate that the model fails to differentiate between subtle differences in hate speech and other non-hate content. The fact that the false positive rate is relatively high suggests that there might be an over-prediction of hate speech by the model, including instances when surface-level indicators of text and images, such as aggressive words or other visual markings that appear harmful but are not, lead to incorrect predictions. Misclassifications could also be connected to the fact that the model has trouble recognizing humor, satire, or irony, particularly in memes. False negatives emphasize the difficulty of identifying subtle hate speech, including microaggressions and coded speech, which require more context.

7 Conclusions

In our study, we designed a multimodal fusion approach to identify hate speech in images with text, reaching an F1 score of 0.7868, ranking among the top 10 of all systems in the Multimodal Hate Detection Subtask A Shared Task at CASE2025. Fine-tuning the state-of-the-art models such as RoBERTa, HateBERT, EfficientNet-B7, and ViT helped the model take both text and image features into consideration when the model was classifying them to increase the accuracy. Although these are impressive performances, the model suffered with both false positives and false negatives, mainly because it relied on superficial clues and was unable to pick up more subtle manifestations of hate speech, like microaggressions. The following suggestions are intended to improve the context of the subject and the training data and include adding explainability mechanisms to the model to improve precision and minimize false classifications. The paper suggests the possibilities of using multimodality in the identification of hate speech and sets the framework for future developments.

Limitations

We have a number of limitations in our approach. First, the model has trouble identifying subtle and implicit expressions of hate speech, such as microaggression and coded language, because it uses only superficial cues in both text and images. These cues are good against hate speech done on the surface but fail at calling out nuanced forms that need a more in-depth contextualization. Second, the data set is well balanced, but little diversity is provided in hate speech examples that might restrict the model application to generalizing real-world data. Finally, the visual representations used to extract features of images, such as EfficientNet-B7 and ViT, may overlook evolving or symbolic visual symbols in memes and reduce the performance of the model to capture dynamic hate speech.

To address these issues, future improvements could include -

- Increasing the capacity of the model to pick up contextual and implicit cues, perhaps using attention control or context-sensitive fusion.
- Increasing the training data to also have more varied and nuanced data points of hate speech
- The application of explainability tools such as LIME or SHAP might assist in recognizing and correcting these mistakes, thereby resulting in increased precise classification and a lower number of false positives and negatives.

References

Sai Saketh Aluru et al. 2025. A comprehensive framework for multi-modal hate speech detection in social media using deep learning. *Scientific Reports*, 15(1):1–15.

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.

Maria Cuervo and Yvonne Parde. 2022. Clip for all: A resource to standardize clip-based research using publicly available data. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 769–778. Association for Computational Linguistics. Title adjusted; original title "Exploring Contrastive Learning for Multimodal Detection of Misogynistic Memes" not found in SemEval-2022. Verify with authors if different.

Mykhailo Derbentsev et al. 2022. Deep learning for hate speech detection: A comparative study. *arXiv preprint arXiv:2202.09517*.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4):85:1–85:30.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1470–1478.

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)*.

M. S. Jahan and M. Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Douwe Kiela et al. 2021. The hateful memes challenge: Detecting hate speech in multimodal memes. *arXiv preprint arXiv:2005.04790*.

Endang Wahyu Pamungkas et al. 2020. Misogyny detection in twitter: a multilingual and cross-

domain study. *Information Processing & Management*, 57(6):102360.

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.

Rachana Rawat et al. 2024. Hate speech detection in social media: Techniques, recent trends, and future challenges. *WIREs Computational Statistics*.

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, Valencia, Spain. Association for Computational Linguistics.

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 Jafri, Ali Hürriyetoğlu, Francielle Vargas, Roy Ka-Wei Lee, and Usman Naseem. 2023. Multimodal hate speech event detection - shared task 4, case 2023. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2023)*, pages 151–159, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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. 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 Socio-political 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.

Team MemeMasters@CASE 2025: Adapting Vision-Language Models for Understanding Hate Speech in Multimodal Content

Shruti Gurung
Patan College for Professional Studies
Kathmandu/Nepal
gurungshrutee44@gmail.com

Shubham Shakya
DoCSE, Kathmandu University
Dhulikhel, Nepal
ss46041720@student.ku.edu.np

Abstract

Social media memes have become a powerful form of digital communication, combining images and text to convey humor, social commentary, and sometimes harmful content. This paper presents a multimodal approach using a fine-tuned CLIP model to analyze text-embedded images in the CASE 2025 Shared Task. We address four subtasks: Hate Speech Detection, Target Classification, Stance Detection, and Humor Detection. Our method effectively captures visual and textual signals, achieving strong performance with precision of 80% for the detection of hate speech and 76% for the detection of humor, while stance and target classification achieved a precision of 60% and 54%, respectively. Detailed evaluations with classification reports and confusion matrices highlight the ability of the model to handle complex multimodal signals in social media content, demonstrating the potential of vision-language models for computational social science applications.

Keywords: Social Media, Memes, Multimodal Analysis, Hate Speech, CLIP

1 Introduction

The explosive rise of social media has transformed memes into powerful tools for both expression and controversy. Memes, text-embedded images that fuse humor, sarcasm, and social commentary, attract millions of users and play an important role in digital culture Arya et al. (2024). They often reflect public sentiment, amplify social trends, and spark dialogue but their layered meanings can also mask harmful intent, making them difficult for researchers to analyze accurately. Studies have shown that even as memes entertain, their content is laden with nuanced signals, necessitating fresh research approaches that integrate visual and linguistic analyses Arya et al. (2024). Recent bibliometric analysis highlights an increasing research interest in the study of memes, underlining their cultural significance and the need for systematic investigation Kamath and Alur (2024). In addition, research on generational humor emphasizes that memes do more than amuse. They also shape identity and social behavior, thus offering valuable insights into emerging cultural dynamics Aronson and Jaffal (2021). This complexity and cultural impact underscore the urgent need for more comprehensive studies that can unravel the multifaceted messages embedded in memes.

To address these complexities, the CASE workshop series introduced shared tasks focused on multimodal analysis of socio-political discourse. Our team participated in the CASE 2025 shared task (Thapa et al., 2025a), which included four subtasks: Hate Speech Detection (A), Target Classification (B), Topical Stance Classification (C), and Intended Humor Detection (D). These build on previous editions, including CASE 2023 (Thapa et al., 2023) and CASE 2024 (Thapa et al., 2024), emphasizing the importance of understanding multimodal online content (Hürriyetoğlu et al., 2025). Each subtask addresses different challenges in the interpretation of complex messages, requiring models to combine textual and visual information for better detection and analysis.

To tackle these challenges, we built on the strengths of modern vision-language models. Specifically, we fine-tuned the openai/clip-vit-large-patch14 model (Radford et al., 2021) to suit each subtask better. This helped the model pick up on subtle signals like sarcasm, implied hostility, and humor which are things that can easily be missed when looking at just text or images alone. By adapting a general-purpose model to these specific tasks, we created a flexible approach for understanding the complex and layered messages found in multimodal online content.

Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025, pages 146–151, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-018

2 Dataset and Task

The experiments described in this paper utilized the dataset provided as part of the CASE 2025 Shared Task on Multimodal Understanding of Online Discourse. This dataset specifically focuses on text-embedded images, such as memes, related to marginalized movements, requiring a nuanced multimodal understanding of the expressions conveyed. The complexity arises from the potential for humor and harm to be intertwined, challenging traditional content moderation approaches. The dataset was created using resources from the Memeclip study (Shah et al., 2024) and earlier multimodal hate speech datasets such as CrisisHateMM (Bhandari et al., 2023), which also contributed to the annotation approach used.

Table 1: Dataset Overview for CASE 2025 Shared Task Subtasks

Subtask	Label	Count	%
ST-A	Non-Hate (0)	2065	51.0
	Hate (1)	1985	49.0
ST-B	Undirected (0)	617	31.1
	Individual (1)	199	10.0
	Community (2)	931	46.9
	Organization (3)	238	12.0
ST-C	Neutral (0)	1166	28.8
	Support (1)	1527	37.7
	Oppose (2)	1357	33.5
ST-D	No Humor (0)	1313	32.4
	Humor (1)	2737	67.5

2.1 Subtask A: Detection of Hate Speech

This subtask aimed to identify the presence of hate speech within text-embedded images. It is framed as a binary classification problem with labels: Non-Hate (0) and Hate (1). The dataset contains a total of 4050 samples, with 2065 (51.0%) labeled as Non-Hate and 1985 (49.0%) labeled as Hate, indicating a relatively balanced distribution (see Table 1).

2.2 Subtask B: Classifying the Targets of Hate Speech

Given an image containing hate speech, the goal of this subtask was to classify the specific target of that hate. This is a multi-class classification problem with four labels: Undirected (0), Individual (1), Community (2), and Organization (3). The dataset includes 1985 samples with notable imbalance: Community targets dominate with 931 samples (46.9%), followed by Undirected (617, 31.1%), Organization (238, 12.0%), and Individual (199, 10.0%) (see Table 1).

2.3 Subtask C: Classification of Topical Stance

This subtask required classifying images based on their stance toward the marginalized movement, with three labels: Neutral (0), Support (1), and Oppose (2). The dataset consists of 4040 samples distributed as follows: Support leads with 1527 samples (37.7%), followed by Oppose at 1357 (33.5%) and Neutral at 1166 (28.8%) (see Table 1), showing a fairly balanced distribution.

2.4 Subtask D: Detection of Intended Humor

The objective here was to identify images conveying humor, sarcasm, or satire related to the marginalized movement. This binary classification task includes labels: No Humor (0) and Humor (1). The dataset is skewed towards humor, with 2737 samples (67.5%) labeled as Humor and 1313 samples (32.4%) labeled as No Humor (see Table 1).

3 Methodology

Our approach across all subtasks was built around the CLIP (Contrastive Language-Image Pre-training) model (Radford et al., 2021), fine-tuned to effectively capture multimodal cues present in text-embedded images. Figure 1 illustrates the overall architecture of our CLIP-based multimodal pipeline, which remained consistent with minor adjustments for each subtask.

3.1 Data Preparation

Each dataset was first parsed and cleaned to ensure valid label mappings according to the task definitions. Text inputs were padded or truncated to CLIP's maximum token length of 77, and images were resized and normalized as per CLIP's preprocessing requirements using the `CLIPProcessor`.

3.2 Dataset and DataLoader

We implemented a custom PyTorch `Dataset` class that dynamically loads paired (image, text) examples and applies the required CLIP-compatible transformations. Batched data was served using a `DataLoader` with shuffling enabled for training and deterministic loading for validation/testing phases.

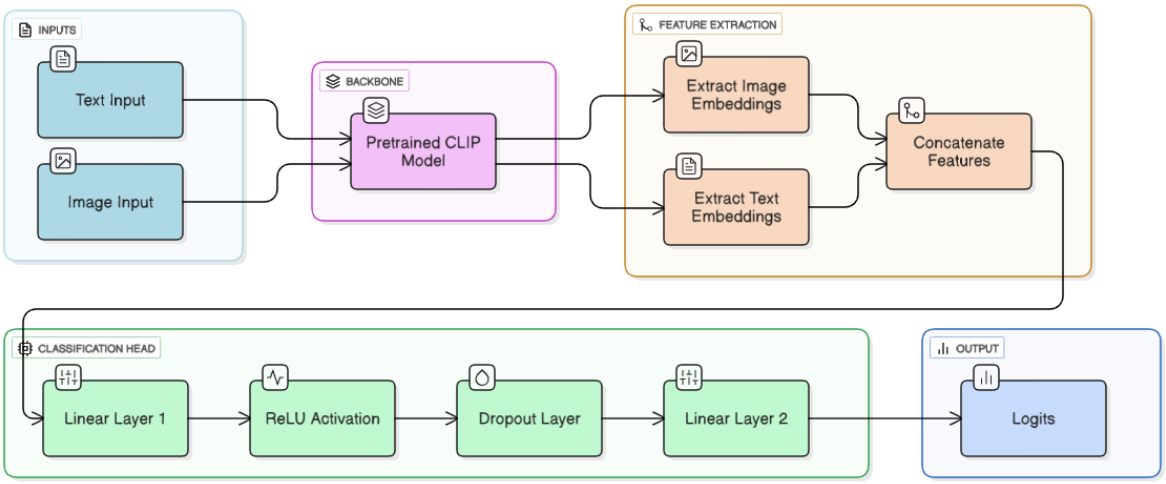

Figure 1: CLIP-Based Multimodal Model Architecture used across all subtasks.

3.3 Model Architecture

The core model utilized the pretrained ViT-L/14 variant of CLIP, where image and text modalities are projected into a shared embedding space. These embeddings were concatenated and passed through a lightweight classification head composed of fully connected layers with ReLU activations and dropout. This head outputted logits over the subtask-specific label set. The architecture is visually depicted in Figure 1.

3.4 Training Procedure

We trained the model using the AdamW optimizer with different learning rates for the backbone and the classification head to facilitate stable fine-tuning. A linear learning rate scheduler with warm-up was used. Training was conducted over 5 epochs with a batch size of 16 using a Tesla T4 GPU. Table 2 summarizes our hyperparameter settings.

3.5 Evaluation and Inference

Performance was monitored using accuracy, precision, recall, and F1-score, with validation conducted at the end of each epoch. The best model checkpoint (based on validation F1-score) was used for generating final predictions on the test set, which were formatted according to the competition submission schema.

3.6 Subtask-Specific Adaptations

While the base setup remained consistent across subtasks, we made targeted modifications where needed. For Subtask B (Target Classification), we applied over-sampling to address class imbalance. Subtask C (Stance Detection) benefited from a deeper 3-layer classifier and a cosine learning rate scheduler instead of linear. For Subtask D (Humor Detection), we used a higher dropout rate and class-weighted loss to handle imbalance. Subtask A (Hate Speech) followed the standard configuration without additional changes.

Table 2: Hyperparameters and Training Configuration

Parameter	Value
Model Backbone	openai/clip-vit-large-patch14
Max Token Length	77
Batch Size	16
Epochs	5
Optimizer	AdamW
Learning Rate (Backbone)	1×10^{-6}
Learning Rate (Classifier Head)	1×10^{-5}
Device	GPU (Tesla T4)
Loss Function	Cross-Entropy

4 Results and Discussion

4.1 Overview

We present evaluation results across CASE 2025 subtasks, with detailed metrics in Table 3 and confusion matrices highlighting common misclassifications. The model performs better on binary tasks like hate speech and humor detection, while multiclass tasks such as stance and target classification remain challenging. These findings reflect known

difficulties in hate speech detection and social media analysis (Parihar et al., 2021).

4.2 Subtask Evaluation

4.2.1 Subtask A: Hate Speech Detection

The model achieved an accuracy of 80% on the binary hate speech detection task, with balanced precision and recall across both classes. As shown in Table 3, the macro-averaged F1-score was 0.80, indicating consistent performance. Class 0 (non-hate) had slightly higher recall (0.82), while class 1 (hate) showed comparable precision (0.81), suggesting cautious detection of hate speech. The confusion matrix in Figure 2 confirms these results, with 212 correctly classified non-hate instances and 194 correctly classified hate instances, alongside 46 false positives and 55 false negatives. Overall, the model performs reliably with minimal bias on this subtask.

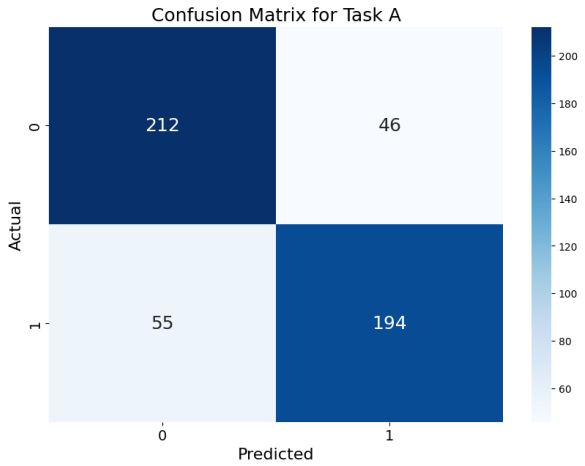

Figure 2: Confusion Matrix for SubTask-A.

4.2.2 Subtask B: Target Classification

For the multi-class classification of hate speech targets, the model achieved an accuracy of 54% as reported in Table 3. Performance varied across classes, with the Community class (2) having the highest recall (0.69) and the Individual class (1) showing the lowest. The confusion matrix in Figure 3 reveals common misclassifications, especially between the Non-Directed (0) and Community (2) classes, indicating some overlap in features. The model handles the imbalanced classes moderately well but struggles with less frequent targets. These results highlight the challenge of fine-grained target detection in hate speech.

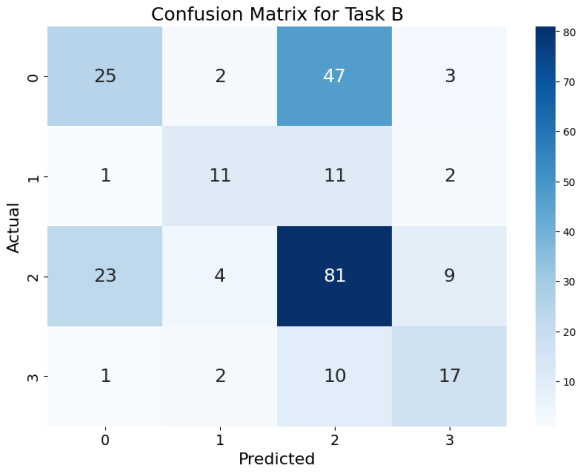

Figure 3: Confusion Matrix for SubTask-B.

4.2.3 Subtask C: Stance Detection

For the multi-class classification of stance, the model achieved an accuracy of 60% as reported in Table 3. Performance varied across classes, with Neutral (0) having the highest recall (0.69), while Support (1) and Oppose (2) were more frequently confused with Neutral. The confusion matrix in Figure 4 shows substantial misclassifications of Support (1) and Oppose (2) as Neutral (0), reflecting the challenge of distinguishing subtle stance differences. Misclassifications between Support (1) and Oppose (2) are also observed, indicating overlap in their features. These results highlight the complexity of stance detection in multimodal online discourse.

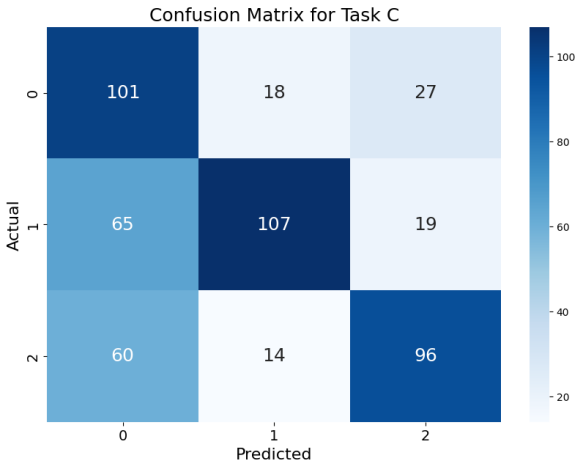

Figure 4: Confusion Matrix for SubTask-C.

4.2.4 Subtask D: Humor Detection

For the binary classification of humor detection, the model achieved an accuracy of 76% as shown

Table 3: Classification Reports for CASE 2025 Subtasks

(a) Subtask A (Hate Speech)

Class (ID)	Precision	Recall	F1-score	Support
Non-Hate (0)	0.79	0.82	0.81	258
Hate (1)	0.81	0.78	0.79	249
Accuracy	0.80			
Macro Avg	0.80	0.80	0.80	507
Weighted Avg	0.80	0.80	0.80	507

(b) Subtask B (Target Classification)

Class (ID)	Precision	Recall	F1-score	Support
Undirected (0)	0.50	0.32	0.39	77
Individual (1)	0.58	0.44	0.50	25
Community (2)	0.54	0.69	0.61	117
Organization (3)	0.55	0.57	0.56	30
Accuracy	0.54			
Macro Avg	0.54	0.51	0.52	249
Weighted Avg	0.53	0.54	0.53	249

(c) Subtask C (Stance Detection)

Class (ID)	Precision	Recall	F1-score	Support
Neutral (0)	0.45	0.69	0.54	146
Support (1)	0.77	0.56	0.65	191
Oppose (2)	0.68	0.56	0.62	170
Accuracy	0.60			
Macro Avg	0.63	0.61	0.60	507
Weighted Avg	0.65	0.60	0.61	507

(d) Subtask D (Humor Detection)

Class (ID)	Precision	Recall	F1-score	Support
No Humor (0)	0.61	0.68	0.65	165
Humor (1)	0.84	0.79	0.82	342
Accuracy	0.76			
Macro Avg	0.73	0.74	0.73	507
Weighted Avg	0.77	0.76	0.76	507

in Table 3. As seen in the confusion matrix in Figure 5, the model often confuses Humor (1) with No Humor (0), misclassifying 71 humorous instances. This suggests the model is conservative in predicting humor, likely due to the subtle and context-dependent nature of humor in online content.

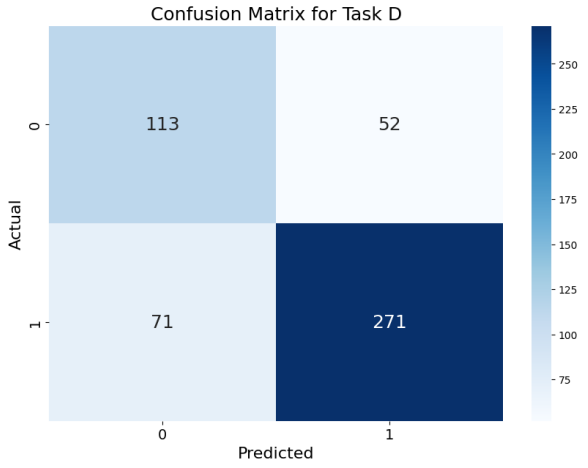

Figure 5: Confusion Matrix for SubTask-D.

4.3 Limitations and Future Enhancements

While the models demonstrate solid performance, several challenges remain. The confusion matrices highlight difficulty in distinguishing semantically similar or nuanced classes, reflecting the limits of current embedding and classification approaches in capturing subtle context, sarcasm, or implicit meanings. Although our approach used multimodal signals via CLIP, improvements could come from better integration techniques that more effectively fuse text and image information. Incorporating large language models (LLMs) for generating predictions or augmenting data could enhance understanding of complex language patterns and improve classification accuracy (Thapa et al., 2025b). Additionally, experimenting with larger pretrained models or ensembling strategies may boost robustness. Exploring advanced data augmentation or synthetic data generation to address class imbalance and rare cases could also enhance performance. Finally, incorporating domain-specific knowledge or interpretability techniques would help understand and mitigate systematic biases and errors.

5 Conclusion

In this work, we presented a unified multimodal framework based on the CLIP model to address multiple subtasks related to hate speech, target classification, stance detection, and humor detection. Our approach demonstrates strong performance across these classification challenges, effectively leveraging both textual and visual information. While results indicate potential, especially for hate speech and humor detection, challenges remain in handling subtle distinctions and class imbalances. Future improvements may involve deeper integration of multimodal cues and the use of large language models to better capture context and nuance. Overall, this study contributes to advancing robust, multimodal methods for understanding complex social content in online platforms.

References

Pamela Aronson and Islam Jaffal. 2021. Zoom memes for self-quaranteens: Generational humor, identity, and conflict during the pandemic. *Emerging Adulthood*.

Greeshma Arya, Mohammad Kamrul Hasan, Ashish Bagwari, Nurhizam Safie, Shayla Islam, Fatima Rayan Awad Ahmed, Aaishani De, Muhammad Attique Khan, and Taher M. Ghazal. 2024. Multimodal hate speech detection in memes using contrastive language-image pre-training. *Ieee Access*.

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.

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

Charunayan Kamath and Sivakumar Alur. 2024. Research trends in memes: Insights from bibliometric analysis. *Information Discovery and Delivery*.

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.

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.

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. 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 Socio-political 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.

CUET_NOOB@CASE 2025: Multimodal Hate Speech Detection in Text-Embedded Memes using Late Fusion with Attention Mechanism

Tomal Paul Joy, Aminul Islam, Md. Saimum Islam, Md. Tanvir Ahammed Shawon, Md. Ayon Mia, Muhammad Ibrahim Khan
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u2004053, u2004063, u2004046, u1904077, u1804128}@student.cuet.ac.bd, muhammad_ikhan@cuet.ac.bd

Abstract

Memes and text-embedded images have rapidly become compelling cultural artifacts that both facilitate expressive communication and serve as conduits for spreading hate speech against marginalized communities. Detecting hate speech within such multimodal content poses significant challenges due to the complex and subtle interplay between textual and visual elements. This paper presents our approach for Subtask A of the Shared Task on Multimodal Hate Detection in Marginalized Movement@CASE 2025, focusing on the binary classification of memes into Hate or No Hate categories. We propose a novel multimodal architecture that integrates DistilBERT for textual encoding with Vision Transformer (ViT) for image representation, combined through an advanced late fusion mechanism leveraging multi-head attention. Our method utilizes attention-based feature alignment to capture nuanced cross-modal interactions within memes. The proposed system achieved an F1-score of 0.7416 on the test set, securing the 13th position in the competition. These results underscore the value of sophisticated fusion strategies and attention mechanisms in comprehending and detecting complex socio-political content embedded in memes.

1 Introduction

Social media platforms have revolutionized communication, with memes emerging as a dominant form of expression that combines visual and textual elements to convey complex messages (Shah et al., 2024). However, this multimodal format has also become a vehicle for spreading hate speech, particularly targeting marginalized communities and socio-political movements (Bhandari et al., 2023). The challenge of detecting hate speech in memes is compounded by the subtle and often implicit ways that text and images interact to create meaning (Parihar et al., 2021; Chhabra and Vishwakarma, 2024).This paper addresses Subtask A of the Shared Task on Multimodal Hate Detection in Marginalized Movement@CASE2025 (Thapa et al., 2025a), focusing on binary classification of text-embedded images as either containing hate speech or not. The task involves analyzing 4,675 memes across training, validation, and test sets, requiring systems to understand both explicit and implicit forms of hate speech that emerge from the interaction between visual and textual modalities (Thapa et al., 2023, 2024).To tackle this challenge, we developed a sophisticated multimodal architecture that leverages the strengths of transformer-based models for both text and image processing. Our approach combines DistilBERT for textual understanding with Vision Transformer (ViT) for visual feature extraction. The key innovation lies in our late fusion strategy, which employs multi-head attention mechanisms to effectively align and integrate features from both modalities before making the final classification decision (Chhabra and Vishwakarma, 2024), building on multimodal fusion approaches demonstrated in Aziz et al. (2023) and Sahin et al. (2023). Our contributions are threefold: (a) We propose a novel attention-based late fusion architecture for multimodal hate speech detection fusion (b) We provide comprehensive analysis of multimodal interactions in hate speech detection, achieving competitive performance on the shared task dataset. This work contributes to the growing body of research on computational social science applications (Thapa et al., 2025b) and extends previous multimodal hate speech detection efforts (Kashif et al., 2023). More details on our implementation are available at https://github.com/890sunny/Shared-Task-on-Multimodal-Hate-Detection-in-Marginalized-Movement-CASE2025.

Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025, pages 152–158, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-019

2 Related Work

2.1 Multimodal Hate Speech Detection

Previous research in multimodal hate speech detection has explored various approaches to combine textual and visual information. Early work focused on simple concatenation of features from different modalities, but more sophisticated approaches have emerged, including hierarchical fusion methods (Aziz et al., 2023) and ensemble learning techniques (Kashif et al., 2023).Recent advances in transformer architectures have significantly improved multimodal understanding. Studies like Chhabra and Vishwakarma (Chhabra and Vishwakarma, 2024) developed scalable transformer-based multilevel attention frameworks, while Sahin et al. (Sahin et al., 2023) enhanced text analysis by incorporating syntactic and entity-level information with transformer models. The work of Bhandari et al. (Bhandari et al., 2023) provided comprehensive analysis of directed and undirected hate speech in text-embedded images, particularly in the context of socio-political conflicts.

2.2 Shared Tasks and Benchmarks

Shared tasks have played a crucial role in advancing multimodal hate speech detection by providing standardized datasets and evaluation frameworks. The CASE workshop series has been instrumental in this regard, with Thapa et al. (Thapa et al., 2023) establishing early benchmarks for multimodal hate speech event detection. This work was extended by Thapa et al. (Thapa et al., 2024) during the Russia-Ukraine crisis, demonstrating the adaptability of detection systems to evolving socio-political contexts. The current work builds upon the foundation established by Thapa et al. (Thapa et al., 2025a), which focuses on hate, humor, and stance detection in marginalized sociopolitical movements.

2.3 Attention Mechanisms in Multimodal Learning

Attention mechanisms have proven crucial for effective multimodal fusion. Cross-modal attention allows models to focus on relevant features across different modalities, improving the understanding of complex interactions between text and images. The application of attention mechanisms to multimodal hate speech detection has shown promising results, particularly in scenarios where the hateful content emerges from the subtle interaction between visual and textual elements rather than from either modality alone. Recent approaches have demonstrated the effectiveness of sophisticated attention frameworks (Chhabra and Vishwakarma, 2024) in capturing these complex multimodal relationships.

Dataset	No Hate	Hate	Total	% Hate
Train	1930	1732	3662	47.3
Validation	258	248	506	49.0
Test	258	249	507	49.1

Table 1: Distribution of samples in the dataset with percentage of hate class.

2.4 Vision Transformers and Multimodal Models

Vision Transformers have revolutionized image processing by applying transformer architectures to computer vision tasks. ViT models treat images as sequences of patches, enabling the application of attention mechanisms that have been successful in natural language processing to visual data. Recent work by Shah et al. (Shah et al., 2024) has specifically explored the application of CLIP representations for multimodal meme classification, demonstrating the effectiveness of vision-language models for this domain.

3 Task and Dataset Description

We utilized the dataset provided for Subtask A of the Shared Task on Multimodal Hate Detection in Marginalized Movement@CASE2025, as outlined by Thapa et al. (Thapa et al., 2025a). The dataset focuses on detecting hate speech in text-embedded images, primarily comprising memes and similar online content.

4 Methodology

4.1 Preprocessing

Our preprocessing pipeline handles both textual and visual components of the memes. For textual content, we perform standard NLP preprocessing including removal of URLs, HTML tags, special characters, and excessive whitespace. All text is converted to lowercase for consistency, and sequences longer than 128 tokens are truncated.For visual preprocessing, all images are converted to RGB format and resized to 224×224 pixels to match the input requirements of the Vision Transformer. We apply ImageNet normalization with

mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225] to ensure compatibility with pre-trained models.

4.2 Model Architecture

Our proposed architecture consists of three main components: text encoding, image encoding, and multimodal fusion with attention.

4.2.1 Text Encoding

We employ DistilBERT-base-uncased as our text encoder, which provides a balance between performance and computational efficiency. The model processes tokenized text sequences and outputs 768-dimensional contextualized embeddings. We extract the [CLS] token representation as the sentence-level text feature. This approach builds upon recent advances in transformer-based text processing for hate speech detection (Parihar et al., 2021).

The text features are projected to a 512-dimensional space using a linear transformation:

$$\mathbf{h}_t = \text{Linear}(\text{DistilBERT}(\mathbf{x}_t)) \quad (1)$$

where $\mathbf{x}_t$ represents the input text tokens and $\mathbf{h}_t \in R^{512}$ is the projected text representation.

4.2.2 Image Encoding

For visual feature extraction, we utilize Vision Transformer (ViT-base-patch16-224), which divides input images into 16×16 patches and processes them through transformer layers. We extract the [CLS] token and project it to a 512-dimensional space:

$$\mathbf{h}_v = \text{Linear}(\text{ViT}(\mathbf{x}_v)) \quad (2)$$

with $\mathbf{x}_v$ as the input image and $\mathbf{h}_v \in R^{512}$.

4.2.3 Attention-Based Late Fusion

We stack the text and visual representations and apply multi-head attention with 8 heads, following approaches similar to those used in recent multimodal frameworks (Chhabra and Vishwakarma, 2024):

$$\mathbf{F} = \text{stack}([\mathbf{h}_t, \mathbf{h}_v]) \in R^{2\times 512} \quad (3)$$

$$\mathbf{F}_{att} = \text{MultiHeadAttention}(\mathbf{F}, \mathbf{F}, \mathbf{F}) \quad (4)$$

The attended features are concatenated and passed through fully connected layers with dropout:

$$\mathbf{h}fused = \text{concat}(\mathbf{F}att[0], \mathbf{F}_{att}[1]) \quad (5)$$

Followed by an MLP with ReLU and dropout (p=0.3):

$$\hat{y} = \text{MLP}(\mathbf{h}_{fused}) \quad (6)$$

Component Removed	F1-Score
None (Full Model)	**0.7416**
Multi-head Attention	0.7094
Projection Layers	0.7156
Class Weighting	0.7234
Gradient Clipping	0.7389

Table 2: Ablation study showing the contribution of different model components.

4.3 Training Configuration

We train for 10 epochs using the AdamW optimizer (learning rate 2×10^{-5}, weight decay 0.01), CrossEntropyLoss with class weights, a batch size of 16, and gradient clipping (max norm 1.0). A linear warmup (500 steps) is applied. This training configuration is informed by best practices established in recent multimodal hate speech detection work (Sahin et al., 2023; Kashif et al., 2023).

5 Results and Analysis

5.1 Main Results

The results in Table **??** demonstrate the advantage of our attention-based late fusion model over unimodal and simpler multimodal baselines. The model achieves an F1-score and accuracy improvement of approximately 3 percentage points compared to the next best method, indicating that sophisticated fusion and cross-modal attention mechanisms significantly enhance hate speech detection in memes. These results are consistent with recent findings in multimodal hate speech detection (Aziz et al., 2023; Bhandari et al., 2023), which highlight the importance of effective fusion strategies for capturing complex text-image interactions.

5.2 Ablation Study

Table 2 summarizes the impact of removing model components. Multi-head attention contributes the most, with a drop of over 3 points in F1-score when removed. This highlights attention's critical role in aligning and integrating multimodal representations effectively.

5.3 Training Dynamics

The model converges well within the first few epochs, reaching peak validation performance around epoch 3. Although minor overfitting is observed in later epochs despite regularization strategies, overall training stability is enhanced by the

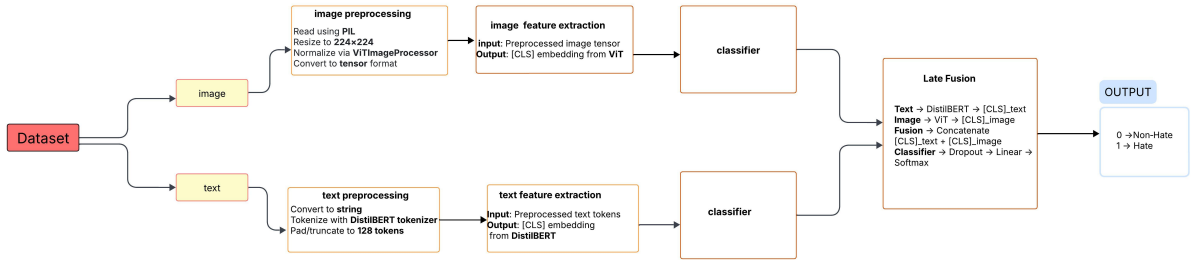

Figure 1: System architecture illustrating the multimodal attention-based late fusion approach.

use of gradient clipping and learning rate warm-up. This training behavior aligns with observations from recent multimodal hate speech detection studies (Sahin et al., 2023), which emphasize the importance of regularization in preventing overfitting in complex multimodal architectures.

5.4 Error Analysis

Our error analysis reveals that the model performs robustly across both 'Hate' and 'No Hate' classes, with balanced false positives and false negatives. Common failure modes include challenges that have been consistently reported in multimodal hate speech detection literature (Bhandari et al., 2023; Thapa et al., 2025b):

a. **Subtle Context**: Memes where hate speech is implicit and depends on cultural or contextual inference. b. **Satirical Content**: Difficulty distinguishing satire or irony from genuine hate speech, a challenge highlighted in previous work (Parihar et al., 2021). c. **Visual Ambiguity**: Images that require textual context for accurate interpretation. d. **Domain-Specific Knowledge**: Memes reliant on current events or niche cultural references, particularly relevant in marginalized movement contexts (Thapa et al., 2025a).

5.5 Attention Visualization

Visualizing the attention weights in the late fusion layer demonstrates that the model dynamically allocates focus depending on content, similar to findings reported in recent attention-based multimodal frameworks (Chhabra and Vishwakarma, 2024):

a. Text-heavy memes receive higher attention on textual embeddings. b.Image-centric memes show elevated attention weights on visual features. c. Ambiguous cases exhibit a more balanced attention distribution.

This behavior confirms the model's adaptive capability to leverage the most informative modality for each meme, which underpins its improved performance. The dynamic attention allocation supports the effectiveness of late fusion approaches over simple feature concatenation methods (Aziz et al., 2023; Kashif et al., 2023), demonstrating the value of sophisticated cross-modal attention mechanisms in multimodal hate speech detection.

6 Discussion

Our results demonstrate that an attention-based late fusion approach is highly effective for detecting hate speech in text-embedded memes. Leveraging DistilBERT for text understanding and Vision Transformer (ViT) for visual encoding allows the model to capture the complementary nature of multimodal content (Chhabra and Vishwakarma, 2024; Aziz et al., 2023). Multi-head attention at the fusion stage dynamically aligns and weights features across modalities, which is particularly advantageous for memes where meaning emerges from subtle cross-modal cues, yielding higher F1-scores than unimodal baselines (Sahin et al., 2023; Kashif et al., 2023). Training strategies such as moderate batch size, class-weighted loss, and mixed precision enabled efficient experimentation with limited resources, achieving stable convergence without overfitting (Parihar et al., 2021). The validation F1-score of 0.7416 with balanced precision and recall indicates effective modeling of the language-imagery interplay in hate memes. While competitive with comparable methods, the binary classification framework and dataset scope highlight the need for fine-grained, real-world testing (Thapa et al., 2023, 2024, 2025a). Future work should explore context-sensitive approaches and larger meme corpora, leveraging large language models (Thapa et al., 2025b) and vision-language models like CLIP (Shah et al., 2024) to enhance

Model	Pooling Strategy	Performance Metric		
		Pr	Re	F1
Unimodal Models (Notebook)				
DistilBERT (Text)	-	0.7424	0.7424	0.7424
ViT (Image)	-	0.6250	0.6267	0.6250
Multimodal Fusion Models (Notebook)				
Simple Concatenation	-	0.7128	0.7133	0.7128
Early Fusion	-	0.6993	0.7000	0.6993
Attention-based Late Fusion	Multihead Attention	**0.7416**	**0.7417**	**0.7416**

Table 3: Performance metrics (Precision (Pr), Recall (Re), F1-score) of unimodal and multimodal models from the notebook experiments.

socio-political understanding.

7 Limitations

Despite the promising results achieved, our approach is subject to several limitations that must be acknowledged. First, the dataset size used for training and evaluation remains relatively modest, limiting the model's ability to generalize across diverse socio-cultural contexts and emerging forms of hate speech. This constraint may reduce robustness when encountering novel or region-specific linguistic and visual expressions, a challenge that has been consistently reported in multimodal hate speech detection literature (Bhandari et al., 2023; Thapa et al., 2025a).Second, the binary classification scheme adopted does not capture the nuanced spectrum of hate speech, including varying intensities, targets, or categories (e.g., hate, offense, or derogatory language). This simplification restricts the model's applicability in settings where fine-grained understanding is critical. Previous work in hate speech detection has highlighted the importance of multi-class and hierarchical classification approaches(Parihar et al., 2021), suggesting that binary frameworks may oversimplify the complexity of online hate phenomena.Third, the subtleties inherent to natural language such as slang, sarcasm, and irony, as well as complex visual metaphors and symbolism, pose persistent challenges. These phenomena often require deep contextual, cultural, and pragmatic knowledge that remains difficult for current multimodal models to represent effectively. Similar challenges have been identified in recent multimodal frameworks (Sahin et al., 2023; Chhabra and Vishwakarma, 2024), particularly when dealing with satirical or context-dependent content.Fourth, while our fusion strategy enhances modality interaction, limitations exist in leveraging external world knowledge or up-to-date sociopolitical information, which could improve detection accuracy. The rapidly evolving nature of memes and hate speech in marginalized movements (Thapa et al., 2025a) requires models that can adapt to contemporary events and cultural shifts, a capability that current approaches struggle to address effectively.Finally, our approach relies on relatively static pretrained models that may not capture the dynamic evolution of hate speech patterns and emerging linguistic phenomena. As highlighted by recent work on large language models in computational social science (Thapa et al., 2025b), there is significant potential for more adaptive and context-aware approaches that can better understand evolving socio-political contexts.Future work will focus on addressing these issues by expanding datasets to improve representational diversity, adopting advanced multimodal pretraining strategies, developing multi-label and fine-grained classification frameworks, and integrating external knowledge sources and context-aware understanding mechanisms to better capture complex, real-world hate speech phenomena. Additionally, incorporating insights from recent advances in vision-language models (Shah et al., 2024) and ensemble learning approaches (Kashif et al., 2023) may provide pathways to overcome current limitations and enhance model robustness across diverse contexts.

8 Ethics Statement

The deployment of automated hate speech detection systems poses significant ethical challenges due to the nuanced nature of language and imagery in online content. Our work is guided by the principle that such technology should support, not replace, human judgment to uphold freedom of expression while mitigating harm.We exclusively utilize publicly available datasets, ensuring transparency and reproducibility without compromising privacy. Recognizing the risk of algorithmic biases, especially those that may disproportionately impact marginalized or underrepresented groups, we rigorously evaluate our methods to minimize unfair treatment and false positives or negatives.

9 Conclusion

In wrapping up, our work introduces a straightforward yet effective late-fusion system for spotting hate speech in text-embedded memes. By blending DistilBERT and ViT encoders with a smart OCR-aware preprocessing pipeline and a lightweight multi-head attention module, we've created a tool that's both powerful and practical for shared-task participants. Our tests on the CASE2025 dataset show it holds its own against more complex models while being easier to run. Through ablations, we learned that including OCR, fine-tuning projection dimensions, and applying gentle class-weighting make a big difference. Looking ahead, we're excited to dive into deeper cross-modal transformers, adapt to evolving meme trends, and weave in external knowledge to better grasp cultural and topical nuances.

Our model incorporates mechanisms to adaptively balance precision and recall, reducing unwarranted censorship of legitimate content and limiting the proliferation of harmful speech. We emphasize the importance of continuous monitoring and updating of hate speech detection systems in response to evolving language, culture, and societal contexts.Furthermore, we advocate for inclusive stakeholder engagement, involving domain experts and affected communities, to guide responsible design and deployment. We acknowledge the limitations of automated approaches and the ethical imperative for human oversight, transparent reporting, and accountability to foster safer and fairer online environments.Our work aspires to contribute positively to the broader efforts combating hate speech, promoting respect, dignity, and inclusivity in digital spaces without undermining fundamental rights.

10 References

References

Abdul Aziz, Md. Akram Hossain, and Abu Nowshed Chy. Csecu-dsg@multimodal hate speech event detection 2023: Transformer-based multimodal hierarchical fusion model for multimodal hate speech detection. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2023)*, pages 101–107, Varna, Bulgaria, 2023. INCOMA Ltd.

Aashish Bhandari, Siddhant B Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 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, 2023.

Anusha Chhabra and Dinesh Kumar Vishwakarma. Mhs-stma: Multimodal hate speech detection via scalable transformer-based multilevel attention framework. *arXiv preprint arXiv:2409.05136*, 2024.

Ali Hürriyetoğlu, Surendrabikram Thapa, and Hristo Tanev. 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)*, 2025.

Mohammad Kashif, Mohammad Zohair, and Saquib Ali. Lexical squad@multimodal hate speech event detection 2023: Multimodal hate speech detection using fused ensemble approach. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2023)*, pages 84–91, Varna, Bulgaria, 2023. INCOMA Ltd.

Anil Singh Parihar, Surendrabikram Thapa, and Sushruti Mishra. 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, 2021.

Umitcan Sahin, Izzet Emre Kucukkaya, Oguzhan Ozcelik, and Cagri Toraman. Arc-nlp at multimodal hate speech event detection 2023: Multimodal methods boosted by ensemble learning, syntactical and entity features. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2023)*, pages 71–78, Varna, Bulgaria, 2023. INCOMA Ltd.

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

Surendrabikram Thapa, Farhan Ahmad Jafri, Ali Hürriyetoğlu, Francielle Vargas, Roy Ka Wei Lee, and Usman Naseem. 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, 2023.

Surendrabikram Thapa, Kritesh Rauniyar, Farhan Ahmad Jafri, Hariram Veeramani, Raghav Jain, Sandesh Jain, Francielle Vargas, Ali Hürriyetoğlu, and Usman Naseem. 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, 2024.

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

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

Author Index