Extended Multimodal Hate Speech Event Detection During Russia-Ukraine Crisis - Shared Task at CASE 2024

Surendrabikram Thapa[1], Kritesh Rauniyar[2], Farhan Ahmad Jafri[3], Hariram Veeramani[4], Raghav Jain[5], Sandesh Jain[1], Francielle Vargas[6], Ali Hürriyetoğlu[7], Usman Naseem[8]

[1]Virginia Tech, USA, [2]Delhi Technological University, India, [3]Jamia Millia Islamia, India, [4]UCLA, USA, [5]University of Manchester, UK, [6]University of São Paulo, Brazil, [7]Wageningen Food Safety Research, Netherlands, [8]Macquarie University, Australia

[1]surendrabikram@vt.edu, [2]rauniyark11@gmail.com, [6]francielleavargas@usp.br

Abstract

Addressing the need for effective hate speech moderation in contemporary digital discourse, the *Multimodal Hate Speech Event Detection* Shared Task made its debut at CASE 2023, co-located with RANLP 2023. Building upon its success, an extended version of the shared task was organized at the CASE workshop in EACL 2024. Similar to the earlier iteration, in this shared task, participants address hate speech detection through two subtasks. *Subtask A* is a binary classification problem, assessing whether text-embedded images contain hate speech. *Subtask B* goes further, demanding the identification of hate speech targets, such as individuals, communities, and organizations within text-embedded images. Performance is evaluated using the macro F1-score metric in both subtasks. With a total of 73 registered participants, the shared task witnessed remarkable achievements, with the best F1-scores in Subtask A and Subtask B reaching **87.27%** and **80.05%**, respectively, surpassing the leaderboard of the previous CASE 2023 shared task. This paper provides a comprehensive overview of the performance of seven teams that submitted results for Subtask A and five teams for Subtask B.

1 Introduction

The constant increase of radicalism and hate around the world has become an urgent global problem. Nowadays, social media has been explored by different radicalism groups to spread hate and terrorism using different data modalities (e.g. text, image, video). In this scenario, the investigation of Hate Speech Detection (HSD) technologies is undoubtedly important since the proposition of automated systems has implications for safe and unprejudiced societies (Vargas et al., 2023).

Nevertheless, there is a wide range of challenges to the detection of multimodal hate speech events on social media, including inaccurate definitions for offensiveness and hate speech, lack of contextual information, and scarce consideration of their social and stereotype bias.

Although there is no consensus related to the definition of hateful and offensive content, most relevant literature distinguishes offensive content and hate speech detection. Offensive content is defined as text, image, or video that disrespects, insults, or attacks the reader containing any form of untargeted profanity (Zampieri et al., 2019). On the other hand, hate speech is defined as a special form of offensive language that attacks or diminishes and incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, or others, and it may occur with different linguistic styles, even in subtle forms as humor and sarcasm (Fortuna and Nunes, 2018). In addition, hate speech is also defined as a particular form of offensive language considering stereotypes to express an ideology of hate (Warner and Hirschberg, 2012).

Given the complex nature of hate speech, it is important to find novel technologies that can aid in the automated detection of hate speech (Parihar et al., 2021). Hate speech detection and moderation via automated techniques become even more complicated when multiple modalities are involved e.g. text and images. In order to bring in new ideas, the shared task on multimodal hate speech detection was organized in CASE 2023 (Thapa et al., 2023). Building on the interests shown by the research community, we have yet again conducted the shared task in CASE 2024.

In this paper, we present a comprehensive overview of the seven registered teams in our extended shared task at CASE 2024. In addition, we describe their proposed approaches, performances, and results, besides the discussion of future advances.The findings of this shared task are ex-

Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), pages 221–228

pected to guide the research direction in finding appropriate research techniques for hate speech and target detection in multimodal settings like text-embedded images.

2 Related Works

Identifying hate speech on social media is an increasingly challenging task that demands the focus of researchers, policy-makers, and society (Jahan and Oussalah, 2023). The majority of studies have mostly concentrated on classifying individual tweets, disregarding the contextual aspects of the discourse (Meng et al., 2023). Various manifestations of hate speech, such as texts, images, and videos, should be identified and addressed swiftly to preserve the decorum of online platforms (Das, 2023). There have been limited attempts to identify text-embedded images for hate speech on social media (Bhandari et al., 2023; Gomez et al., 2020). Text-embedded images are visuals that include text as an integral part of their composition. Text-embedded images are frequently seen in several settings, including online social networks (OSNs) and video content (Das et al., 2023; Chhabra and Vishwakarma, 2023). The image functions as a means of establishing context, while the text that comes with it communicates the information contained throughout that context. Current research on hate speech classification has a main issue which is the lack of structured data creation and diverging annotation schema, resulting in weak adaptability of supervised-learning models to new datasets (Jin et al., 2023). To overcome this problem, Bhandari et al. (2023) proposed a dataset of text-embedded images related to the Russia-Ukraine crisis. Building on the dataset, this shared task aims to bring researchers and professionals to address the problem of hate speech and its target detection in text-embedded images.

3 Dataset

We utilized the same dataset as CASE 2023 (Thapa et al., 2023; Hürriyetoğlu et al., 2023) for our shared task. This dataset, known as CrisisHateMM, was introduced in work by Bhandari et al. (2023) and comprises a collection of 4,723 text-embedded images, all centered around the Russia-Ukraine Crisis (Thapa et al., 2022). Within this dataset, 2,058 images were found to be free from any instances of hate speech, whereas the remaining 2,665 images included elements of hate speech. Among the images containing hate speech, a subset of 2,428 text-embedded images displayed instances of targeted or directed hate speech. For our shared task, we exclusively considered text-embedded images that had directed hate speech, and those that did not have any hate speech. This selection resulted in the use of a total of 4,486 text-embedded images. To ensure a balanced and representative data set, we divide it into distinct training, evaluation, and test sets for Subtasks A and B. This division was carried out in a stratified manner, maintaining a consistent split ratio of approximately 80-10-10, mirroring the approach employed in CASE 2023 (Thapa et al., 2023). The details of the dataset can be found in Table 1.

Subtask	Classes	Train	Eval	Test
Subtask A	Hate	1942	243	243
	No Hate	1658	200	200
Subtask B	Individual	823	102	102
	Community	335	40	42
	Organization	784	102	98

Table 1: Statistics of the dataset at train, evaluation, and test phase of our shared task

4 Shared Task Description

According to Koushik et al. (2019), people from various cultural and educational backgrounds are sharing their thoughts on Twitter, Facebook, and Tumbler, thanks to the abrupt rise in popularity of microblogging services. Their ideas occasionally use language that is harsh, violent, or insulting and target a particular group of individuals who share something in common, such as a gender, an ethnic group, a belief system, or a geographic area. Because hate speech on social media has increased, it is exceedingly time-consuming and costly to manually detect hate speech on these platforms.

4.1 Subtask A: Hate Speech Detection

The objective of this task is to determine the presence of hate speech in text-embedded images. The dataset employed for this subtask comprises annotated images, categorizing them into two labels: 'Hate Speech' and 'No Hate Speech'. The dataset's focus is on images with embedded text, and the annotation process involves identifying whether the content falls into the hate speech category or not. The binary labels, 'Hate Speech' and 'No Hate Speech', precisely characterize the classification

criteria for this task, providing a clear distinction between instances with offensive content and those without offensive content.

4.2 Subtask B: Targets of Hate Speech Detection

The objective of this specific task is to classify the specific targets of hate speech within text-embedded images. These images, containing hateful text, encompass a range of potential targets having diverse categories. However, our subtask specifically concentrates on identifying three predefined targets as specified in the dataset used for our shared task. The annotated targets in the dataset include 'community', 'individual', and 'organization'. As a result, our primary goal is to accurately pinpoint and categorize these particular targets within the text-embedded images that exhibit hate speech. This task involves understanding and classifying the hateful content, focusing on recognizing whether it is directed toward a community, an individual, or an organization. The aim is to enhance understanding and identification of hate speech by observing these predetermined target categories within the context of text-embedded images.

5 Evaluation and Competition

This section explains the nature of our competition, including the system for calculating rankings and other important details.

5.1 Evaluation Metrics

We employed accuracy, precision, recall, and macro F1-score to evaluate the performance of the participants' contributions. The macro F1-score sorting method was used to establish the participants' rank.

5.2 Competition Setup

We used the Codalab[1] to organize our competition. There were two stages to the competition: an evaluation stage where participants were introduced to the Codalab system, and a testing phase where the ultimate leaderboard ranking was established based on performance.

Registration: For our competition, 73 individuals registered in total. It was evident from the wide variety of email domains that were utilized that the competition was effective in drawing participants from different parts of the world. 7 teams out of the total number of registrants sent in their predicted outcomes.

Competition Timelines: On November 1, 2023, training and evaluation data were made available, marking the beginning of the competition. The first phase was the evaluation phase. Participant familiarization with Codalab was the primary goal of the evaluation phase, therefore participants were also given access to the evaluation data labels. Then, on November 30, 2023, test data without any ground truth labels were released, indicating the beginning of the test phase. The test period was extended until January 7, 2024, in response to requests from several participants, from its original end date of January 5, 2024. The system description paper submission deadline was ultimately decided upon as January 16, 2024.

6 Participants' Methods

In this section, we describe the various methods used by the participants who submitted the system description paper.

6.1 Overview

A total of 7 participants submitted scores for subtask A, while 5 participants submitted to subtask B. The leaderboards for subtask A and subtask B are presented in Table 2 and Table 3, respectively. In both subtasks, CLTL achieved the top performance, surpassing the other models by a significant margin. These models also outperformed the highest scores achieved by ARC-NLP in the same shared task, which was conducted during CASE 2023 at RANLP 2023. In the subsequent subsections, we provide detailed system descriptions for each participating team.

6.2 Methods

Below, we provide a summary of the system descriptions provided by the participating teams in the shared task. These summaries are derived from the approaches detailed by the participants in their system description papers.

6.2.1 Subtask A

CLTL (Wang and Markov, 2024) proposed a method that includes separate text and image processing modules coupled with a simple MLP and softmax, providing an optimal alternative to Large

[1]The competition page can be found here: https://codalab.lisn.upsaclay.fr/competitions/16203.

Rank	Team Name	Codalab Username	Accuracy	Precision	Recall	F1-score
1	CLTL (Wang and Markov, 2024)	Yestin	**87.36**	**87.20**	**87.37**	**87.27**
2	MasonPerplexity (Gangul et al., 2024)	Sadiya_Puspo	83.52	83.47	83.78	83.47
3	AAST-NLP (El-Sayed and Nasr, 2024)	AhmedElSayed	76.98	76.76	76.76	76.76
4	YYama (Yamagishi, 2024)	YYama	75.85	75.88	76.13	75.80
5	CUET_Binary_Hackers	Asrarul_Hoque_Eusha	68.62	68.61	68.79	68.55
6	-	kriti7	46.05	46.45	46.44	46.05
7	Team +1	pakapro	49.66	56.83	53.23	44.08

Table 2: Sub-task A (Hate Speech Classification) Leaderboard, Ranked by Macro F1-Score. All scores are presented as percentages (%). The highest score in each column is highlighted in bold. 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.

Vision Language Models (LVLMs). This method increases design flexibility and analytic capability. The presentation is distinguished by its cleanliness, straightforward but original ideas, and clarity. The results show that the implementation stands out as a competitive benchmark. It shows how multimodal models need not always be trained together for a specific task and a modular approach with simple MLP-based feature fusion could work at the same level if not better. This could also be easily noticed with some of the authors (Yamagishi, 2024) who used a pre-trained LVLM and achieved considerably lower scores than the one proposed in (Wang and Markov, 2024). This could also point toward the significance of fine-tuning in LVLM optimization. Overall, the approach exhibits a simple yet effective pipeline for hate speech detection in image-based data. Their approach achieved the first position with performances noted in Table 2.

MasonPerplexity (Gangul et al., 2024) experimented with various models like BERTweet-large (Ushio and Camacho-Collados, 2021; Ushio et al., 2022), BERT-base (Devlin et al., 2019), XLM-R (Conneau et al., 2020a), and GPT-3.5 [2] in their implementation. The test F1-score of the models were 75%, 81%, and 83% for BERT-base, BERTweet-large, and XLM-R respectively. GPT models also showed remarkable performance with a F1-score of 82% in the test dataset for fine-tuned GPT 3.5.

AAST-NLP (El-Sayed and Nasr, 2024) initially fine-tuned the bert variants, RoBERTa (Liu et al., 2019), XLM-RoBERTa (Conneau et al., 2020b), and HateBERT (Caselli et al., 2021) on all of the datasets to attain the best results. They then proposed the **top-k** ensemble technique and various multimodal models, such as ViT Dosovitskiy et al. (2021) model and Swin Liu et al. (2021) as feature extractor to achieve higher macro F1-score. In order to get the highest F1-score, they utilized the 'Top-3' ensemble strategy, which combined several BERT versions. They have employed the most recent CLIP (Contrastive Language–Image Pre-training) Radford et al. (2021) model, which combines textual and visual data via cross-fusion and concatenation. 85.40% was the greatest recall on CLIP (Concat), and 85.50% and 85.44% were the highest precision and F1-score, respectively, on the **Top-3** ensemble technique. Out of 7 teams, they were able to secure the third position in this task.

YYama (Yamagishi, 2024) proposed an approach whose goal was to optimize user prompts for the LLaVa-1.5B LVLM architecture by applying simple prompt engineering approaches for hate-speech detection. Although there have been other LVLM-based techniques for image-based hate-speech recognition in recent years (Hermida and Santos, 2023; Van and Wu, 2023). Therefore the methodology is not fully novel; the author offers insightful information at the prompt level. The study indicates that simple prompts tend to perform better than complicated ones. This difference in performance is attributed to a narrower filter that is used to identify difficult instructions inside the prompts. The author makes strong arguments and highlights how the model uses a variety of implicit meanings for 'no hate speech' to effectively handle open-ended queries. On the other hand, adding more definitions causes the internal definition set to shrink, which might increase the number of false negatives. Overall, the paper presented us with an approachable method deploying existing LVLM models for specified tasks with open-ended and simpler prompts, which, contrary to popular methods such as chain-of-thoughts, presents us with a

[2] https://platform.openai.com/docs/models

lower barrier to generating appropriate responses. Their approach attained the fourth position with performances noted in Table 2.

6.2.2 Subtask B

CLTL (Wang and Markov, 2024) employ the same foundational model for subtask A, with only the output layer undergoing modification. Despite minimal customization, their approach surpasses all others and establishes a new benchmark. The key to their success lies in the embedded features captured and fused by the MLP. This layer effectively represents all essential features related to hate speech, simplifying the MLP's task in discerning whether the hate is directed towards an organization, individual, or community. This results in an impressive over 18% improvement over the baseline and a 2-5% lead over the previous state-of-the-art models. Furthermore, the paper underscores the importance and significance of fine-tuning in achieving these remarkable results. Lastly, the strategic use of RoBERTa, particularly in conjunction with Twitter's social interaction data, provides the authors with significant prior knowledge of the competition's domain, contributing significantly to their success. Their approach attained the first position in subtask B with performances noted in Table 3.

AAST-NLP (El-Sayed and Nasr, 2024) first optimized RoBERTa (Liu et al., 2019), XLM-RoBERTa (Conneau et al., 2020b), and HateBERT (Caselli et al., 2021) models of BERT (Devlin et al., 2019) variations on all datasets in order to get the greatest performance. To obtain a better score, they conducted experiments utilizing the **top-k** ensemble technique and the latest CLIP (Contrastive Language–Image Pre-training) model, which integrates textual and visual input through cross-fusion and concatenation. They used the 'Top-3' ensemble technique, combining multiple BERT variants, to obtain the greatest F1-score possible. Using the **Top-3** ensemble approach, they were able to achieve the maximum values of all three metrics: precision, recall, and F1-score, which were 74.99%, 82.73%, and 77.03%, respectively. In a challenge of five teams in this subtask, they took second place.

MasonPerplexity (Gangul et al., 2024) used the ensemble of BERTweet-large (Ushio and Camacho-Collados, 2021; Ushio et al., 2022), BERT-base (Devlin et al., 2019), and XLM-R (Conneau et al., 2020a) in order to achieve their best score. They also tested with various standalone models like BERTweet-large (Ushio and Camacho-Collados, 2021; Ushio et al., 2022), BERT-base (Devlin et al., 2019), XLM-R (Conneau et al., 2020a), and GPT 3.5. With the ensemble model, the F1-score was 67%. Similarly, the standalone models performed 61%, 64%, and 66% with BERT-base, XLM-R, and BERTweet-large respectively. Similarly, with various configurations of GPT, the authors achieved F1-scores of 53%, 57%, and 63% with zero shots, few shots, and fine-tuned settings, respectively.

7 Discussion

The results and methods presented in this shared task demonstrate diverse approaches to hate speech classification, shedding light on the complexity of addressing this pressing issue. CLTL's modular approach (Wang and Markov, 2024), separating text and image processing, exemplifies the adaptability of multimodal models. MasonPerplexity's exploration of various language models underscores the importance of thoughtful model selection (Gangul et al., 2024), while AAST-NLP's ensemble technique and CLIP utilization highlight the benefits of combining multiple models and modalities (El-Sayed and Nasr, 2024). YYama's focus on prompt optimization provides an accessible method for deploying existing models with straightforward prompts (Yamagishi, 2024). These approaches collectively contribute to the ongoing advancements in hate speech detection, emphasizing the significance of both model architecture and prompt design. The healthy competition and diversity of strategies among the participating teams contribute to the ongoing progress in the field of hate speech research.

8 Conclusion

In conclusion, the Multimodal Hate Speech Event Detection Shared Task, first introduced at CASE 2023 and extended to CASE 2024, provided a platform for exploring innovative approaches to combat hate speech in contemporary digital discourse. This shared task witnessed significant participation from a total of 73 registered participants, resulting in remarkable achievements in both Subtask A and Subtask B. The top-performing models in Subtask A achieved an impressive F1-score of 87.27%, while Subtask B saw a top F1-score of 80.05%, surpassing the previous CASE 2023 shared task

Rank	Team Name	Codalab Username	Accuracy	Precision	Recall	F1-score
1	CLTL (Wang and Markov, 2024)	Yestin	**82.64**	**81.48**	**79.07**	**80.05**
2	AAST-NLP (El-Sayed and Nasr, 2024)	AhmedElSayed	80.99	82.73	74.99	77.03
3	MasonPerplexity (Gangul et al., 2024)	Sadiya_Puspo	71.49	67.59	67.27	67.41
4	CUET_Binary_Hackers	Asrarul_Hoque_Eusha	51.24	34.50	41.35	37.48
5	Team +1	pakapro	28.10	28.12	30.31	24.78

Table 3: Sub-task B (Targets of Hate Speech Classification) Leaderboard, Ranked by Macro F1-score. All scores are presented as percentages (%). The highest score in each column is highlighted in bold. 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.

leaderboard. The diverse methods employed by the participating teams, including modular multimodal models, careful model selection, ensemble techniques, and prompt optimization, highlight the various approaches to tackle the complex problem of hate speech detection. These efforts collectively contribute to advancing the field and emphasize the importance of continuous research in addressing this critical issue in online discourse. The shared task fosters healthy competition and encourages future research in hate speech detection and multimodal analysis.

Acknowledgements

We would like to acknowledge Diego Alves, Samuel Guimarães, Isabelle Carvalho, and Siddhant Bikram Shah for helping us provide detailed reviews for the shared task. Their insights were instrumental in shaping the feedback provided to the participants. Moreover, this work is supported by the European Research Council Politus Project (ID:101082050) and European Union's HORIZON projects EFRA (ID: 101093026) and ECO-Ready (ID: 101084201).

Broader Impact

The Multimodal Hate Speech Event Detection Shared Task has the potential to profoundly impact society by advancing the development of more accurate and effective hate speech detection models. These advancements can create safer online spaces, reduce the spread of hate speech, and foster constructive digital discourse. However, ethical considerations are paramount, as the deployment of automated detection systems must balance the imperative to combat hate speech with concerns about potential biases and limitations that may inadvertently suppress free expression or disproportionately target specific groups. Additionally, from a technological perspective, this shared task drives innovation in multimodal AI research, benefiting fields beyond hate speech detection, such as content moderation, multimedia analysis, and human-computer interaction. Furthermore, in academia, it enriches the study of hate speech detection by providing benchmark datasets and promoting collaboration among researchers, leading to a deeper understanding of the challenges involved and the development of novel methodologies.

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 1993–2002.

Tommaso Caselli, Valerio Basile, Mitrovic Jelena, Granitzer Michael, et al. 2021. Hatebert: Retraining bert for abusive language detection in english. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25. Association for Computational Linguistics.

Anusha Chhabra and Dinesh Kumar Vishwakarma. 2023. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, pages 1–28.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Unsupervised cross-lingual representation learning at scale.

Mithun Das. 2023. Classification of different participating entities in the rise of hateful content in social media. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 1212–1213.

Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. 2023. Hatemm: A multi-modal dataset for hate video classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1014–1023.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Ahmed El-Sayed and Omar Nasr. 2024. AAST-NLP at Multimodal Hate Speech Event Detection 2024 : A Multimodal Approach for Classification of Text-Embedded Images Based on CLIP and BERT-Based Models. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Malta. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

Amrita Gangul, Al Nahian Bin Emran, Sadiya Sayara Chowdhury Puspo, Md Nishat Raihan, Dhiman Goswami, and Marcos Zampieri. 2024. Mason-Perplexity at Multimodal Hate Speech Event Detection 2024: Hate Speech and Target Detection Using Transformer Ensembles. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Malta. Association for Computational Linguistics.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1470–1478.

Paulo Cezar de Q Hermida and Eulanda M dos Santos. 2023. Detecting hate speech in memes: a review. *Artificial Intelligence Review*, pages 1–19.

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.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, page 126232.

Yiping Jin, Leo Wanner, Vishakha Laxman Kadam, and Alexander Shvets. 2023. Towards weakly-supervised hate speech classification across datasets. *arXiv preprint arXiv:2305.02637*.

Garima Koushik, K. Rajeswari, and Suresh Kannan Muthusamy. 2019. Automated hate speech detection on twitter. In *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, pages 1–4.

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*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows.

Qing Meng, Tharun Suresh, Roy Ka-Wei Lee, and Tanmoy Chakraborty. 2023. Predicting hate intensity of twitter conversation threads. *Knowledge-Based Systems*, page 110644.

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.

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, Aditya Shah, Farhan Jafri, Usman Naseem, and Imran Razzak. 2022. A multimodal dataset for hate speech detection on social media: Case-study of russia-Ukraine conflict. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 1–6, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Asahi Ushio, Francesco Barbieri, Vitor Sousa, Leonardo Neves, and Jose Camacho-Collados. 2022. Named entity recognition in Twitter: A dataset and analysis on short-term temporal shifts. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 309–319, Online only. Association for Computational Linguistics.

Asahi Ushio and Jose Camacho-Collados. 2021. T-NER: An all-round python library for transformer-based named entity recognition. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 53–62, Online. Association for Computational Linguistics.

Minh-Hao Van and Xintao Wu. 2023. Detecting and correcting hate speech in multimodal memes with large visual language model. *arXiv preprint arXiv:2311.06737*.

Francielle Vargas, Isabelle Carvalho, Ali Hürriyetoğlu, Thiago Pardo, and Fabrício Benevenuto. 2023. Socially responsible hate speech detection: Can classifiers reflect social stereotypes? In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1187–1196, Varna, Bulgaria.

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)*, Malta. Association for Computational Linguistics.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada. Association for Computational Linguistics.

Yosuke Yamagishi. 2024. YYama@Multimodal Hate Speech Event Detection 2024: Simpler Prompts, Better Results - Enhancing Zero-shot Detection with a Large Multimodal Model. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Malta. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. 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 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.