Multimodal Hate Speech Event Detection - Shared Task 4, CASE 2023

Surendrabikram Thapa
Department of Computer Science, Virginia Polytechnic Institute and State University, United States of America
sbt@vt.edu

Farhan Ahmad Jafri
Department of Computer Science, Jamia Millia Islamia, India
farhanjafri88888@gmail.com

Ali Hürriyetoğlu
KNAW Humanities Cluster DHLab, The Netherlands
ali.hurriyetoglu@dh.huc.knaw.nl

Francielle Vargas
Institute of Mathematical and Computer Sciences, University of São Paulo, Brazil
francielleavargas@usp.br

Roy Ka-Wei Lee
Singapore University of Technology and Design Singapore, Singapore
roy_lee@sutd.edu.sg

Usman Naseem
College of Science and Engineering, James Cook University, Australia
usman.naseem@jcu.edu.au

Abstract

Ensuring the moderation of hate speech and its targets emerges as a critical imperative within contemporary digital discourse. To facilitate this imperative, the shared task **Multimodal Hate Speech Event Detection** was organized in the sixth CASE workshop co-located at RANLP 2023. The shared task has two sub-tasks. The sub-task A required participants to pose hate speech detection as a binary problem i.e. they had to detect if the given text-embedded image had hate or not. Similarly, sub-task B required participants to identify the targets of the hate speech namely individual, community, and organization targets in text-embedded images. For both sub-tasks, the participants were ranked on the basis of the F1-score. The best F1-score in sub-task A and sub-task B were 85.65 and 76.34 respectively. This paper provides a comprehensive overview of the performance of 13 teams that submitted the results in Subtask A and 10 teams in Subtask B.

1 Introduction

The rise of social media has altered the global communication and information landscape, allowing people from all walks of life to share their opinions and perspectives on a wide range of topics, including heated geopolitical events (Overbey et al., 2017; Chen and Zimbra, 2010). This free-flowing exchange of ideas, however, has not been without difficulties. The rapid proliferation of hate speech, which includes harsh language, disrespectful statements, and discriminatory rhetoric directed at individuals or groups based on their ethnicity, nationality, or beliefs, is one of the most alarming concerns afflicting online platforms (Parihar et al., 2021). In times of political crisis, such as the Russia-Ukraine Crisis, the prevalence of hate speech becomes even more pronounced (Thapa et al., 2022). Its impact goes beyond dividing communities; it also brings about considerable concerns for sustaining peace and stability in regions facing conflict-related issues.

Text-embedded images have gained popularity due to their easy sharability and the combination of visual and textual elements, making them a common mode for information sharing (Chen et al., 2022; Bhandari et al., 2023; Lee et al., 2021). However, this convenience also has a downside – it amplifies the prevalence of hate speech in social media. To combat the propagation of hate content through text-embedded images, the identification of hate speech within such media holds significant importance (Cao et al., 2022; Pramanick et al., 2021b; Sharma et al., 2022). By detecting and curbing hate speech within these images, we can work towards maintaining a healthier digital environment. In an attempt to curb hate speech in the context of the Russia-Ukraine crisis, Bhandari et al. (2023) proposed a multimodal dataset of 4,723 text-embedded images annotated for presence of hate speech, direction of hate speech (targeted vs untargeted) and targets of hate speech. Building on this groundwork and to attract greater attention toward the issue of hate speech in text-embedded images, we introduced a shared task at the CASE 2023 workshop (co-located with RANLP 2023) utilizing the dataset. The shared task has two subtasks: subtask

Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text, pages 151–159, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-089-2_020

A which deals with the identification of hate speech and subtask B which deals with the identification of targets in hate speech. Through this shared task, we intend to stimulate active engagement and collaboration in addressing this critical challenge of identifying and mitigating hate speech within the digital landscape, specifically in the context of text-embedded images.

The rest of the paper is organized as follows: Section 2 gives a brief outlook of the related works in multimodal hate speech classification. In section 3, the subtasks of the shared task are presented. Similarly, section 4 describes the CrisisHateMM dataset in brief. Section 5 describes the system that we used in the competition along with the evaluation metrics. Similarly, section 6 sheds light on the methodologies used by the teams that submitted the system description papers. Section 7 gives a brief analysis of the system descriptions, and section 8 finally concludes the paper.

2 Related Work

The task of detecting hate speech in social media has gained significant traction, primarily focusing on text-based content (Alam et al., 2022; Chhabra and Vishwakarma, 2023). However, there has been lesser efforts in classification of text-embedded images for hate speech in social media (Gomez et al., 2020; Bhandari et al., 2023). In recent times, there has been a notable surge in scholarly interest towards identifying hate speech in memes or images containing text (Ji et al., 2023; Hermida and Santos, 2023; Karim et al., 2022; Yang et al., 2022, 2019a; Perifanos and Goutsos, 2021). Memes often combine images and text with the intention of humor. On the other hand, text-embedded images are essentially images that incorporate text within them. This category encompasses not only memes but also other forms of textual-visual content, such as screenshots taken from TV headlines. In these cases, the image itself serves to provide context, while the accompanying text conveys the information within that context. While meme analysis has been a focal point for researchers, the examination of hate speech in these text-embedded images deserves equal attention. The introduction of this shared task stems from the recognition of this research gap.

Similarly, the exploration of memes or multimodal textual-visual data has predominantly concentrated on the broader scope of general social media platforms. The efforts to create dedicated datasets and conduct research within specific contexts have been quite limited. Recently, some research have shown efforts to understand such multimodal textual-visual data for specific contexts and applications. For instance, Pramanick et al. (2021a) investigated harmful memes and their targets in the context of the COVID-19 pandemic. They labeled COVID-19-related memes to indicate harmfulness and the targets of these harmful memes. Expanding on this work, Pramanick et al. (2021b) also studied memes related to the US election using the same labeling approach. Additionally, Naseem et al. (2023) introduced a dataset containing 10,244 memes critical of vaccines. These initiatives are gradually paving the way for future research that aligns with specific contexts. This shared task is also an attempt to attract the attention of the research community, encouraging their involvement in context-oriented investigations.

3 Task Description

According to Warner and Hirschberg (2012), hate speech is a particular form of offensive language that considers stereotypes to express an ideology of hate. Here, we assume that offensive language is a type of opinion-based information that is highly confrontational, rude, or aggressive (Zampieri et al., 2019), which may be led explicitly or implicitly (Vargas et al., 2021; Poletto et al., 2021). In the same settings, hate speech is a particular form of offensive language used against target groups, mostly based on their social identities.

3.1 Subtask 1: Hate Speech Detection

The goal of this task is to identify whether the given text-embedded image contains hate speech or not. The dataset used for this subtask consists of text-embedded images, and these images are annotated to indicate the presence or absence of hate speech. More precisely, the dataset for this sub-task comprises two labels: “Hate Speech” and “No Hate Speech”.

3.2 Subtask 2: Identification of Targets of Hate Speech

The goal of this subtask is to identify the targets of hate speech in a given hateful text-embedded image. Although hate speech text-embedded images may contain various potential targets falling into numerous categories, our subtask focuses solely

on identifying three predetermined targets outlined within the dataset used in our shared task. The text-embedded images in the dataset are annotated for "community", "individual" and "organization" targets. Consequently, our objective centers on the identification of these particular targets within text-embedded images featuring hate speech.

4 Dataset

In our shared task, we used the CrisisHateMM dataset (Bhandari et al., 2023). This dataset consists of a total of 4,723 text-embedded images centered around the Russia-Ukraine Crisis (Thapa et al., 2022). Within these 4,723 text-embedded images, 2,058 did not have any instances of hate speech, while the remaining 2,665 contained elements of hate speech. Among these 2,665 images with hate speech, a subset of 2,428 text-embedded images exhibited instances of targeted or directed hate speech. In our shared task, we used only text-embedded images that exhibited directed hate speech and those that did not have any hate speech. Thus, a total of 4,486 text-embedded images were used in our shared task. We split the dataset into train, evaluation, and test stages for both subtasks A and B in a stratified manner, maintaining a proportionate split ratio of approximately 80-10-10.

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

5 Evaluation and Competition

This section describes our competition environment including ranking methods and other details regarding the competition.

5.1 Evaluation Metrics

In order to assess the performance of participants' submissions, we used accuracy, precision, recall, and macro F1-score. The rank of the participants was determined by sorting based on the macro F1-score.

5.2 Competition Setup

We hosted our competition using the Codalab[1]. The competition had two phases: an evaluation phase, which introduced participants to the Codalab system, and a testing phase which determined the final leaderboard ranking based on performance.

Registration: A total of 51 participants registered for our competition. The diverse range of email domains used indicated that the competition successfully attracted individuals from various geographical regions. Among all the registered participants, a total of 13 teams submitted their predictions.

Competition Timelines: The competition started on May 1, 2023, with the release of training and evaluation data. The first phase was the evaluation phase. As the purpose of the evaluation phase was to make participants familiarize with codalab, the evaluation data labels were also provided to participants. Subsequently, the test phase started on June 15, 2023, with the release of test data that didn't have any ground truth labels. Originally planned to conclude on June 30, 2023, the test phase was extended to July 7, 2023, in response to multiple participant requests. Finally, the deadline for submitting the system description paper was set for July 24, 2023.

6 Participants' Methods

6.1 Overview

A total of 13 participants submitted scores for subtask A, while subtask B received 10 successful submissions. The leaderboards for subtasks A and B are presented in Table 2 and 3 respectively. Notably, in both subtasks, ARC-NLP (Sahin et al., 2023) achieved the highest performance in terms of the F1-score, with scores of 85.65 for subtask A and 76.34 for subtask B. Our next step involves an in-depth discussion of each team's approaches to gain a thorough understanding of the technical intricacies involved.

6.2 Methods

Below, we provide a summary of the systems from the eight teams that submitted description papers, organized based on their leaderboard ranking. Among these submissions, seven papers have

[1]The competition page can be found here: https://codalab.lisn.upsaclay.fr/competitions/13087.

Rank	Team Name	Codalab Username	Accuracy	Precision	Recall	F1-score
1	ARC-NLP (Sahin et al., 2023)	arc-nlp	**85.78**	**85.63**	**85.67**	**85.65**
2	-	bayesiano98	85.33	85.28	85.61	85.28
3	IIC_Team (Singh et al., 2023)	karanpreet_singh	84.65	84.76	85.08	84.63
4	-	DeepBlueAI	83.52	83.35	83.56	83.42
5	CSECU-DSG (Aziz et al., 2023)	csecudsg	82.62	82.44	82.52	82.48
6	Ometeotl (Armenta-Segura et al., 2023)	Jesus_Armenta	81.04	80.94	81.21	80.97
7	SSN-NLP-ACE (K et al., 2023)	Avanthika	79.01	78.81	78.78	78.80
8	VerbaVisor (Esackimuthu and Balasundaram, 2023)	Sarika22	78.56	78.49	78.06	78.21
9	-	rabindra.nath	78.33	78.42	77.68	77.88
10	Lexical Squad (Kashif et al., 2023)	md_kashif_20	73.59	73.72	72.7.	72.87
11	GT	lueluelue	52.60	52.19	52.19	52.19
12	Team + 1	pakapro	49.66	49.39	49.38	49.36
13	ML_Ensemblers	Sathvika.V.S	57.79	72.40	53.34	42.94

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.

been accepted for inclusion in the proceedings of the CASE workshop.

6.2.1 Subtask A

ARC-NLP (Sahin et al., 2023) leveraged syntactic features from the text extracted from the dataset along with ensemble learning in order to predict the presence of hate speech. The information from textual and visual encoders is used to train the multi-layer perception (MLP) (Murtagh, 1991). Similarly, XGBoost (Chen and Guestrin, 2016), Light Gradient Boosting Machine (LGBM) (Alzamzami et al., 2020), and Gradient Boosting Machine (GBM) (Natekin and Knoll, 2013; Ayyadevara and Ayyadevara, 2018) are trained on syntactical and Bag of Words-based features (Zhang et al., 2010). A weighted ensemble (Hürriyetoğlu et al., 2022; Sahin et al., 2022) is used to make the final decision. This method stands as the first method with an F1-score of 85.65.

IIC_Team (Singh et al., 2023) implemented XLM-Roberta-base, BiLSTM, XLNet base cased, and ALBERT on the CrisisHateMM (Bhandari et al., 2023) dataset, consisting of over text-embedded images related to the Russia-Ukraine conflict. The models were fine-tuned on the training sets to enhance hate speech identification, in which they slit the dataset in 80% for training and 20% for validation. Lastly, a robust preprocessing step was performed to prepare the textual data. The authors obtained a high performance presenting an impressive F1 score of 84.62 for sub-task 1 using XLM-Roberta-base. Finally, even though in this proposal the authors did not provide any evaluation related to potential social bias in hate speech technologies (Davani et al., 2023; Vargas et al., 2023), for future works, they aim to tackle strategies towards social bias mitigation, as well as improve the amount of data and its diversity in order to obtain more generalized and accurate results.

CSECU-DSG (Aziz et al., 2023) used a multimodal approach by contextualizing text characteristics using the BERT transformers model. The Bi-LSTM was used to understand long-term contextual relationships and facilitate the extraction of hate speech from the text recovered from images. The ViT transformers model was used to extract visual information from photographs. They used a multi-sample dropout method after combining the outputs of the multimodal and BiLSTM modules to arrive at the final prediction. By achieving an F1-score of 82.48 and an accuracy of 82.62, this technique ranked fifth in subtask A.

Ometeotl (Armenta-Segura et al., 2023) used the pre-trained transformer approach BertForSequenceClassification model with the bert-base-uncased architecture from huggingface[2]. They didn't utilize any preprocessing for subtask A and achieved an F1 score of 80.97. The authors secured the 6th rank in subtask A.

SSN-NLP-ACE (K et al., 2023) extracted the text from text-embedded images using Google Vision API and extracted the features using the TF-IDF (Adhikari et al., 2021) approach. They used the traditional machine learning approach i.e. support vector machine (SVM). In the SVM, the closest data points are the support vectors in finding the optimal plane. The kernel applied in SVM is RBF (Radial Basis Function). The authors tuned the parameters to maximize F1-score to 78.80 and an accuracy of 79.01 in subtask A.

[2] https://huggingface.co/

Rank	Team Name	Codalab Username	Accuracy	Precision	Recall	F1-score
1	ARC-NLP (Sahin et al., 2023)	arc-nlp	**79.34**	**76.37**	**76.36**	**76.34**
2	-	bayesiano98	77.27	73.30	75.54	74.10
3	IIC_Team (Singh et al., 2023)	karanpreet_singh	72.31	71.05	68.94	69.73
4	VerbaVisor (Esackimuthu and Balasundaram, 2023)	Sarika22	71.49	68.41	67.77	68.05
5	CSECU-DSG (Aziz et al., 2023)	csecudsg	69.01	65.75	65.25	65.30
6	-	DeepBlueAI	69.83	66.48	64.62	65.25
7	Ometeotl (Armenta-Segura et al., 2023)	Jesus_Armenta	64.05	67.93	56.48	56.77
8	SSN-NLP-ACE (K et al., 2023)	Avanthika	64.05	70.13	53.84	52.58
9	ML_Ensemblers	Sathvika.V.S	52.89	48.88	44.44	43.32
10	pakapro	Team + 1	35.12	35.59	34.42	33.42

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.

VerbaVisor (Esackimuthu and Balasundaram, 2023) implemented Artificial Neural Networks (ANN) (Mishra and Srivastava, 2014) model along with the ALBERT (Lan et al., 2019) model for this subtask. Out of these two ALBERT performed the best with a F1-score of 78.21. The ANN model performed poorly as compared to ALBERT.

Lexical Squad (Kashif et al., 2023) used an approach to combine both textual and visual information from the text-embedded images. They used a combined representation from different unimodal models: XLNet (Yang et al., 2019b) and BERT (Kenton and Toutanova, 2019) for textual features and Inception-V3 (Szegedy et al., 2016) for visual features. Stacking was used to generate a combined representation. This approach gave them a F1-score of 74.96 which is above 3 points improvement when using XLNet alone and above 5 points improvement when using BERT alone. When solely utilizing Inception-V3, they achieved an F1-score of 48.11. The empirical evaluations by the authors showed that the approach yielded poor performances when a model had to leverage a lot of visual information to make decisions.

ML_Ensemblers used a variety of algorithms, which includes Naive Bayes (Rish et al., 2001), k-Nearest Neighbors (KNN) (Jiang et al., 2007), Random Forest (Breiman, 2001), Decision Trees (Kotsiantis, 2013), and Support Vector Machine (SVM) (Cortes and Vapnik, 1995; Pisner and Schnyer, 2020). Among these algorithms, Naive Bayes displayed the highest performance with an F1-score of 42.94. It's important to note that the mentioned approach is not an ensemble, as each algorithm was assessed separately rather than being combined into a unified model. The approach ranked 13th in subtask A.

6.2.2 Subtask B

ARC-NLP (Sahin et al., 2023) made use of entity features along with CLIP (Radford et al., 2021) embeddings to create a feature that was leveraged to classify targets of hate speech. Similar to the approach for subtask A, the ensemble methods were then used to make the final decision. The method was ranked first in the competition with an F1-score of 76.34. The importance of NER in hate speech and target classification has been an interest of the academic community and this method reaffirms that the NER characteristics are very important.

IIC_Team (Singh et al., 2023) implemented XLM-Roberta-base, BiLSTM, XLNet base cased, and ALBERT on the CrisisHateMM (Bhandari et al., 2023) dataset related to the Russia-Ukraine conflict. The authors obtained an F1 score of 69.73 for sub-task 2 using XLM-Roberta-base.

VerbaVisor (Esackimuthu and Balasundaram, 2023) applied ALBERT to approach the problem of target detection in our shared task. They were able to get the fourth rank with an F1-score of 68.05.

CSECU-DSG (Aziz et al., 2023) used the multimodal technique in which they adjusted the BERT (Kenton and Toutanova, 2019) transformers model to extract the text's contextualized properties. The Vision Transformers (ViT) (Dosovitskiy et al., 2020) model was used to extract the visual information from the given image, and the Bi-LSTM was used to learn the long-term contextual dependency that enables the model to extract the hate information present in the context. On top of the outputs from the multimodal and BiLSTM modules, the multi-sample dropout strategy is then applied to obtain the final prediction. This approach gave them an F1-score of 65.30 and an accuracy of 69.01.

Ometeotl (Armenta-Segura et al., 2023) employed the huggingface bert-base-uncased architecture with the pre-trained transformer method BertForSequenceClassification model. Unlike subtask A, for subtask B, they used preprocessing outside of BERT processing of the text, such as eliminating special letters or stopwords, and they received an F1 score of 56.77. The authors placed the seventh rank in subtask B. The case study of different examples led them to hypothesize that image features are more important in target identification than hate speech classification.

SSN-NLP-ACE (K et al., 2023) employed the TF-IDF technique to extract the features from the text of text-embedded images. They approached subtask B using the conventional machine-learning method of Logistic Regression (Nick and Campbell, 2007). It is a technique for statistical analysis that makes use of probability estimates. The hyperparameters were optimized by the authors and an F1-score of 52.58 was achieved.

ML_Ensemblers employed multiple algorithms for target detection. They utilized various algorithms namely Naive Bayes algorithm (Rish et al., 2001; Thapa et al., 2020), k-Nearest Neighbors (kNN) (Jiang et al., 2007), Random Forest (Breiman, 2001), Decision Tree (Kotsiantis, 2013), and Support Vector Machine (SVM) (Cortes and Vapnik, 1995; Pisner and Schnyer, 2020). Among these, the multinomial Naive Bayes algorithm performed the best with an F1-score of 43.32.

7 Discussion

The methods from different participants gave interesting insights into various methods. Particularly, transformer-based methods were seen to be more effective. Most participants utilized BERT-based variations to extract textual features from the dataset. For the extraction of visual features, participants turned to vision transformers, CLIP (Radford et al., 2021), and established methods like Inception-V3. The methodology proposed by Sahin et al. (2023) suggested that syntactical and entity features are equally important to leverage textual information from the dataset, particularly from instances that were related to the identification of targets of hate speech. While it is important to comprehend the utility of transformer-based models, K et al. (2023) suggested that traditional machine learning algorithms can also give a satisfactory performance in hate speech classification. While their algorithm excelled in subtask A, addressing target identification remained challenging for such traditional machine learning approaches. The promising direction for future research is to explore the applications of vision-language models specifically pretrained for the classification of hate speech in text-embedded images of memes.

8 Conclusion

In conclusion, through our shared task at CASE 2024, we were able to contribute to promoting the research and interest in hate speech and target classification in text-embedded images. The shared task was successful in attracting over 50 participants. The participants altogether made over 250 submissions on the test set. The highest performance of F1-score 85.65 was achieved in subtask A and F1-score 76.34 in subtask B. This shows that there is still scope for improvement in the tasks proposed in our shared task. Building on the momentum of this successful shared task, we intend to continue the shared task in the future with more subtasks in languages other than English. This expansion will aim to foster a more inclusive understanding of hate speech detection that goes beyond linguistic and cultural boundaries.

Acknowledgments

The organizers would like to thank Fiona Anting Tan (National University of Singapore) for her technical assistance in setting up the codalab. Additionally, we express our sincere appreciation to the reviewers who provided important feedback on the system description papers.

References

Surabhi Adhikari, Surendrabikram Thapa, Priyanka Singh, Huan Huo, Gnana Bharathy, and Mukesh Prasad. 2021. A comparative study of machine learning and nlp techniques for uses of stop words by patients in diagnosis of alzheimer's disease. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6625–6643.

Fatimah Alzamzami, Mohamad Hoda, and Abdulmotaleb El Saddik. 2020. Light gradient boosting machine for general sentiment classification on short texts: a comparative evaluation. *IEEE access*, 8:101840–101858.

Jesus Armenta-Segura, César Jesús Núñez-Prado, Grigori Olegovich Sidorov, Alexander Gelbukh, and Rodrigo Francisco Román-Godínez. 2023. Ometeotl@Multimodal Hate Speech Event Detection 2023: Hate speech and text-image correlation detection in real life memes using pre-trained bert models over text. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Varna, Bulgaria (Hybrid). Association for Computational Linguistics.

V Kishore Ayyadevara and V Kishore Ayyadevara. 2018. Gradient boosting machine. *Pro machine learning algorithms: A hands-on approach to implementing algorithms in python and R*, pages 117–134.

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)*, Varna, Bulgaria (Hybrid). 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 1993–2002.

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

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.

Hsinchun Chen and David Zimbra. 2010. Ai and opinion mining. *IEEE Intelligent Systems*, 25(3):74–80.

Keyu Chen, Ashley Feng, Rohan Aanegola, Koustuv Saha, Allie Wong, Zach Schwitzky, Roy Ka-Wei Lee, Robin O'Hanlon, Munmun De Choudhury, Frederick L Altice, et al. 2022. Categorizing memes about the ukraine conflict. In *International Conference on Computational Data and Social Networks*, pages 27–38. Springer.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

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

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20:273–297.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Sarika Esackimuthu and Prabavathy Balasundaram. 2023. VerbaVisor@Multimodal Hate Speech Event Detection 2023: Hate speech detection using transformer model. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Varna, Bulgaria (Hybrid). 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, 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.

Junhui Ji, Wei Ren, and Usman Naseem. 2023. Identifying creative harmful memes via prompt based approach. In *Proceedings of the ACM Web Conference 2023*, pages 3868–3872.

Liangxiao Jiang, Zhihua Cai, Dianhong Wang, and Siwei Jiang. 2007. Survey of improving k-nearest-neighbor for classification. In *Fourth international conference on fuzzy systems and knowledge discovery (FSKD 2007)*, volume 1, pages 679–683. IEEE.

Avanthika K, Mrithula KL, and Thenmozhi D. 2023. SSN-NLP-ACE@Multimodal Hate Speech Event Detection 2023: Detection of hate speech and targets using logistic regression and svm. In *Proceedings of the 6th Workshop on Challenges and Applications of*

Automated Extraction of Socio-political Events from Text (CASE), Varna, Bulgaria (Hybrid). Association for Computational Linguistics.

Md Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Md Shajalal, and Bharathi Raja Chakravarthi. 2022. Multimodal hate speech detection from bengali memes and texts. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 293–308. Springer.

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)*, Varna, Bulgaria (Hybrid). Association for Computational Linguistics.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Sotiris B Kotsiantis. 2013. Decision trees: a recent overview. *Artificial Intelligence Review*, 39:261–283.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

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*, pages 5138–5147.

Manish Mishra and Monika Srivastava. 2014. A view of artificial neural network. In *2014 international conference on advances in engineering & technology research (ICAETR-2014)*, pages 1–3. IEEE.

Fionn Murtagh. 1991. Multilayer perceptrons for classification and regression. *Neurocomputing*, 2(5-6):183–197.

Usman Naseem, Jinman Kim, Matloob Khushi, and Adam G Dunn. 2023. A multimodal framework for the identification of vaccine critical memes on twitter. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 706–714.

Alexey Natekin and Alois Knoll. 2013. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7:21.

Todd G Nick and Kathleen M Campbell. 2007. Logistic regression. *Topics in biostatistics*, pages 273–301.

Lucas A Overbey, Scott C Batson, Jamie Lyle, Christopher Williams, Robert Regal, and Lakeisha Williams. 2017. Linking twitter sentiment and event data to monitor public opinion of geopolitical developments and trends. In *Social, Cultural, and Behavioral Modeling: 10th International Conference, SBP-BRiMS 2017, Washington, DC, USA, July 5-8, 2017, Proceedings 10*, pages 223–229. Springer.

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.

Konstantinos Perifanos and Dionysis Goutsos. 2021. Multimodal hate speech detection in greek social media. *Multimodal Technologies and Interaction*, 5(7):34.

Derek A Pisner and David M Schnyer. 2020. Support vector machine. In *Machine learning*, pages 101–121. Elsevier.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(3):477–523.

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.

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.

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.

Irina Rish et al. 2001. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46.

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)*, Varna, Bulgaria (Hybrid). Association for Computational Linguistics.

Umitcan Sahin, Oguzhan Ozcelik, Izzet Emre Kucukkaya, and Cagri Toraman. 2022. Arc-nlp at case 2022 task 1: Ensemble learning for multilingual

protest event detection. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 175–183.

Shivam Sharma, Firoj Alam, Md Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, Tanmoy Chakraborty, et al. 2022. Detecting and understanding harmful memes: A survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 5597–5606.

Karanpreet Singh, Vajratiya Vajrobol, and Nitisha Aggarwal. 2023. IIC_Team@Multimodal Hate Speech Event Detection 2023: Detection of hate speech and targets using xlm-roberta-base. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Varna, Bulgaria (Hybrid). Association for Computational Linguistics.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Surendrabikram Thapa, Aditya Shah, Farhan Ahmad Jafri, Usman Naseem, and Imran Razzak. 2022. A multi-modal dataset for hate speech detection on social media: Case-study of russia-ukraine conflict. In *CASE 2022-5th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text, Proceedings of the Workshop*. Association for Computational Linguistics.

Surendrabikram Thapa, Priyanka Singh, Deepak Kumar Jain, Neha Bharill, Akshansh Gupta, and Mukesh Prasad. 2020. Data-driven approach based on feature selection technique for early diagnosis of alzheimer's disease. In *2020 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE.

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

Francielle Vargas, Fabiana Goes, Isabelle Carvalho, Fabrício Benevenuto, and Thiago Pardo. 2021. Contextual-lexicon approach for abusive language detection. In *Proceedings of the Recent Advances in Natural Language Processing*, pages 1438–1447, Held Online.

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.

Chuanpeng Yang, Fuqing Zhu, Guihua Liu, Jizhong Han, and Songlin Hu. 2022. Multimodal hate speech detection via cross-domain knowledge transfer. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4505–4514.

Fan Yang, Xiaochang Peng, Gargi Ghosh, Reshef Shilon, Hao Ma, Eider Moore, and Goran Predovic. 2019a. Exploring deep multimodal fusion of text and photo for hate speech classification. In *Proceedings of the third workshop on abusive language online*, pages 11–18.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019b. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

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 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1415–1420, Minnesota, United States.

Yin Zhang, Rong Jin, and Zhi-Hua Zhou. 2010. Understanding bag-of-words model: a statistical framework. *International journal of machine learning and cybernetics*, 1:43–52.