IIC_Team@Multimodal Hate Speech Event Detection 2023: Detection of Hate Speech and Targets using Xlm-Roberta-base

Karanpreet Singh
Institute of Informatics and Communication
University of Delhi
Delhi, India
karanpreet.singh@iic.ac.in

Vajratiya Vajrobol
Institute of Informatics and Communication
University of Delhi
Delhi, India
tiya101@south.du.ac.in

Nitisha Aggarwal
Institute of Informatics and Communication
University of Delhi
Delhi, India
nitisha@south.du.ac.in

Abstract

Hate speech has emerged as a pressing issue on social media platforms, fueled by the increasing availability of multimodal data and easy internet access. Addressing this problem requires collaborative efforts from researchers, policymakers, and online platforms. In this study, we investigate the detection of hate speech in multimodal data, comprising text-embedded images, by employing advanced deep learning models. The main objective is to identify effective strategies for hate speech detection and content moderation. We conducted experiments using four state-of-the-art classifiers: XLM-Roberta-base, BiLSTM, XLNet base cased, and ALBERT, on the CrisisHateMM dataset, consisting of over 4700 text-embedded images related to the Russia-Ukraine conflict. The best findings reveal that XLM-Roberta-base exhibits superior performance, outperforming other classifiers across all evaluation metrics, including an impressive F1 score of 84.62 for sub-task 1 and 69.73 for sub-task 2. Additionally, it is worth highlighting that our study achieved the remarkable feat of securing the 3rd position in both sub-tasks. The future scope of this study lies in exploring multimodal approaches to enhance hate speech detection accuracy, integrating ethical considerations to address potential biases, promoting fairness, and safeguarding user rights. Additionally, leveraging larger and more diverse datasets will contribute to developing more robust and generalised hate speech detection solutions.

1 Introduction

Hate speech on social media has become a major issue, with online platforms being used to denigrate and degrade people or entire groups based on their colour, religion, ethnicity, or handicap (Parihar et al., 2021). In the virtual world, the concept of hate speech can be complex and nuanced, making it difficult to address effectively (Mathew et al., 2019; Banks, 2010; Das, 2023). To address the issue, international conventions, and multilateral initiatives have been developed, however, implementing laws in the virtual sphere remains a difficult undertaking. Despite social media firms' efforts, suppressing hate speech is an ongoing process. The increasing amount of multimedia content, powered by quicker and more accessible mobile internet, has altered the social media environment. Instagram, Snapchat, Vine, and TikTok have championed multimedia, prompting established behemoths like Facebook and Twitter to follow suit. This transition has shifted social media from a predominantly text-based environment to one in which video, audio, and photographs take center stage, allowing users to express themselves in more interesting and diverse ways (Castaño-Pulgarín et al., 2021).

The detection of hate speech during political events is especially important for preserving democracy, reducing violence, protecting vulnerable communities, and fostering civil dialogue. Effective detection ensures fair elections, platform integrity, national cohesion, and informed decision-making while balancing free speech protection with actions to eliminate harmful content. Traditional moderation procedures, such as manual text and multimedia inspection, confront substantial restrictions as a result of the massive volume of data created on social media platforms. Human moderators are unable to keep up with the exponential increase in content, resulting in delays in recognizing and correcting hate speech, allowing harmful content to propagate unchallenged. Furthermore, human moderators' biases and subjective interpretations can lead to inconsistent results. To handle the size and pace of data growth, automated hate speech identification systems are crucial.

Automated hate speech detection systems utilise artificial intelligence (AI) techniques to analyse large volumes of data and identify content that con-

Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text, pages 136–143, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-089-2_018

tains hate speech or offensive language these systems rely on Natural Language Processing (NLP) algorithms to preprocess and transform the text data into numerical representations, such as word embeddings. AI models, like recurrent neural networks (RNNs), long short-term memory (LSTM) networks, or transformers, are then employed to extract contextual and semantic features from the text. The model is trained on labeled datasets, learning to recognize patterns and characteristics indicative of hate speech (Smitha et al., 2018; Beskow et al., 2020; De la Pena Sarracén et al., 2018).

To address the need for large datasets to train AI models, the CrisisHateMM Dataset (Bhandari et al., 2023) is introduced, aiding the Shared task on Multimodal Hate Detection at CASE 2023 (Thapa et al., 2023). This dataset includes two primary tasks, with sub-task 1 focusing on classifying text-embedded images into two categories: hate speech and non-hate speech. sub-task 2 involves the classification of targets in the text-embedded images into three categories: individual, organisation, and community. Within the datasets, predominantly comprised of images linked to the Russia-Ukraine conflict, this widespread political event has been the subject of significant hateful language and has been thoroughly examined by researchers (Thapa et al., 2022). By employing a subtask-based methodology, this research approach allows for detailed analysis and interpretability, providing valuable insights into hate speech characteristics and target identification in social media content. The dataset's multimodal and contextually relevant annotations facilitate benchmarking and advancements in combating hate speech on social media platforms.

The paper proposes a novel approach for hate speech identification utilising the textual model Xlm-Roberta-base, achieving impressive results. In sub-task 1, the approach achieves an accuracy of 84.65% and an F1 score of 84.63%. In sub-task 2, it demonstrates solid performance with an accuracy of 72.31% and an F1 score of 69.73%. The method effectively detects hate speech in text-embedded images, showcasing its strong performance in this aspect. Additionally, the study shows significant improvement in target recognition in images with objectionable text, showcasing the efficacy of the proposed approach in enhancing hate speech detection and target identification. Notably, the paper secured the 3rd position in both tasks, signifying its competitive standing within the competition. The accomplishments of this study make a substantial contribution to the field of hate speech identification and further highlight its rank and achievements in the competition.

The paper begins with a concise introduction to the problem of hate speech on social media. It then provides a comprehensive review of previous research on the topic and the technological advancements in recent years, including various approaches, methodologies, datasets, and experimental findings. The dataset is described in detail, statistical analysis is used to gain insights, and preprocessing procedures are covered. The article introduces the approach using NLP models (XLM-Roberta-base, BiLSTM, XLNet base cased, and ALBERT) and reports on how well they perform in identifying hate speech. Results demonstrate the models' efficacy and the discussion analyses the results and discusses constraints. The conclusion highlights the importance of the research in preventing hate speech while summarising the major contributions. References list the sources used to conduct the study.

2 Literature survey

Hate speech on social media is a worrying issue when people utilise online venues to disseminate harmful or discriminatory content, fostering animosity and division. The pervasive effects it has on social cohesiveness, mental health, and actual violence highlight the urgent need for effective content moderation measures to address and reduce this problem.

A study (Gitari et al., 2015) explored the development of a classifier to detect hate speech in web discourses, specifically focusing on race, nationality, and religion themes. They employed sentiment analysis techniques, including subjectivity detection, to identify and rate the polarity of sentiment expressions. By creating a hate speech lexicon based on subjectivity and semantic features, the model effectively classified hate speech. Experimental results with a hate corpus demonstrated the practical applicability of the approach in real-world web discourse scenarios. Researchers (Djuric et al., 2015) also tackled the challenge of hate speech detection in online user comments. Hate speech, defined as abusive speech targeting specific group characteristics like ethnicity, religion, or gender, poses a significant problem for websites that allow user feedback, leading to negative consequences

for their online business and user experience. To address this issue, the paper proposed a novel approach using neural language models to learn distributed low-dimensional representations of comments. These representations are then utilised as inputs to a classification algorithm, effectively addressing the issues of high dimensionality and sparsity that had previously hindered the state-of-the-art hate speech detection methods. As a result, their approach demonstrated high efficiency and effectiveness in detecting hate speech in online comments. The study (MacAvaney et al., 2019) addressed the escalating problem of hate speech dissemination in online content and the difficulties confronted by automatic approaches in detecting such content in text. These challenges encompassed the intricacies of language, divergent definitions of hate speech, and limited availability of data for training and testing these systems. Additionally, the lack of interpretability in many recent approaches posed a significant hurdle, making it arduous to comprehend the rationale behind the system's decisions. To overcome these obstacles, the paper proposed a multi-view Support Vector Machine (SVM) approach, which achieved nearly state-of-the-art performance in hate speech detection while maintaining simplicity and providing more easily interpretable decisions compared to neural methods. The paper concluded by discussing both technical and practical challenges that still persist in this area, emphasising the need for further research to enhance hate speech detection systems for online content.

In 2022, Alkomah et al conducted a comprehensive study on hate speech detection systems, reviewing textual features, machine learning models, and datasets (Alkomah and Ma, 2022). The analysis of 138 relevant papers revealed that many approaches lack consistency in detecting various hate speech categories. The dominant methods often involve combining multiple deep learning models, while several hate speech datasets were found to be small and unreliable for detection tasks. The study provides valuable insights into the complexities of hate speech and highlights the need for improved approaches and larger, more reliable datasets to effectively combat hate speech and foster healthier online communities. Another research in the same year (Rana and Jha, 2022) addressed the pressing need to monitor hate speech on social media platforms, particularly in multimedia content. While text-based filtering has been extensively studied, detecting hate speech in multimedia presents unique challenges. A preliminary study revealed that the speaker's emotional state significantly influences hateful content, prompting the paper to focus on auditory and semantic features. Introducing the first multimodal deep learning framework, the study combines emotional auditory features with semantics to detect hate speech effectively. Results demonstrate improved detection compared to text-based models. Additionally, a new Hate Speech Detection Video Dataset (HSDVD) is introduced, filling the gap in available datasets for this purpose. This research contributes to advancing hate speech detection in multimedia, providing a valuable resource to combat hateful content on social media platforms. (Mazari et al., 2023) conducted a study dedicated to multi-aspect hate speech detection on social media. The overwhelming amount of unfiltered toxic content, including cyberbullying, cyberstalking, and hate speech, has become a significant challenge and a focus of active research. The proposed approach utilises a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model combined with Deep Learning (DL) models to create ensemble learning architectures. The DL models incorporate Bidirectional Long-Short Term Memory (Bi-LSTM) and/or Bidirectional Gated Recurrent Unit (Bi-GRU) on FastText and GloVe word embeddings. Individual training of these models on multi-label hateful datasets and their combination with BERT results in highly effective hate speech detection on social media. By leveraging recent word embedding techniques and DL architectures in conjunction with BERT, the study achieves an impressive ROC-AUC score of 98.63%, significantly enhancing hate speech detection capabilities in multi-aspect scenarios. Recently, one more research by Liam Hebert et al. (2023) introduced the Multi-Modal Discussion Transformer (mDT), a groundbreaking multi-modal graph-based transformer model designed for hate speech detection in online social networks. Unlike traditional text-only methods, mDT takes a holistic approach by considering both text and images when labelling a comment as hate speech. The model leverages graph transformers to capture contextual relationships within the entire discussion surrounding a comment and utilises interwoven fusion layers to combine text and image embeddings, rather than

processing different modalities separately. Comparative evaluations against text-only baselines and extensive ablation studies showcase the superior performance of mDT. The paper concludes by emphasising the significance of multimodal solutions in delivering social value in online contexts and highlights that capturing a holistic view of conversations significantly advances the detection of anti-social behaviour like hate speech. This research presents a promising step towards more effective hate speech detection methods by considering both textual and visual cues in social media discussions (Hebert et al., 2023).

3 Methodology

The study utilised Figure 1 to split the training data into 80% training sets and 20% validation sets. Before model input, a preprocessing step was performed to prepare the textual data. The Xlm-RoBERTa-base, BiLSTM, XLNet base cased, and ALBERT models were fine-tuned on the training sets to enhance hate speech identification. Testing predictions were then generated using the fine-tuned models, demonstrating the effectiveness of the proposed approach in hate speech detection and target identification.

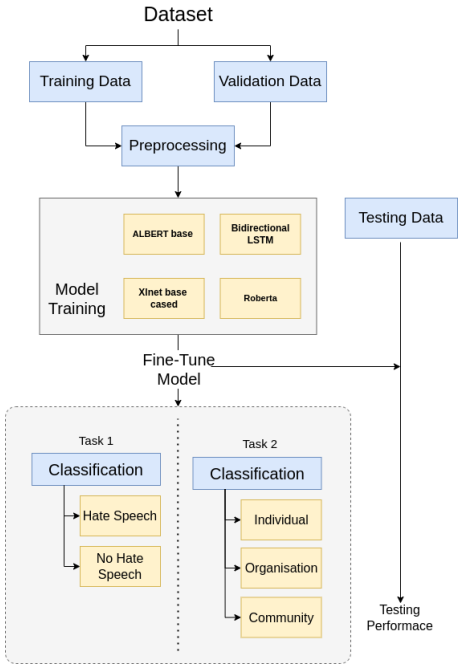

Figure 1: Process Flow of Hate Speech Detection and Target Identification Processes.

Class Name	Number of Words	Number of Unique Words	Maximum Text length (character)	Average Words (per post)
No Hate	82532	14974	1673	49
Hate	63721	12224	1297	32

Table 1: Distribution of extracted text length for 'Hate Speech' and 'Non-Hate speech.

3.1 Dataset

The CrisisHateMM dataset contains over 4700 text-embedded images related to the Russia-Ukraine conflict. It includes 2665 instances of hate speech and 2058 instances of non-hate speech. The dataset is further categorised into directed hate speech (2428 instances) and undirected hate speech (237 instances), with annotations for targets classified into individual (1027), community (417), and organisational (984) categories.

Table 1 indicates regarding sub-task 1 that the "No Hate" class has more posts and a higher number of words compared to the "Hate" class. However, further analysis is necessary to fully understand the significance of these differences. For sub-task 2, target detection is valuable for comprehending the distribution of textual content, post lengths, and word usage across different categories (Individual, Community, and Organisation) in the dataset (Table 2).

Class Name	Number of Words	Number of Unique Words	Max Text Length	Avg Words (per post)
Individual	25.9k	6.4k	1082	31.42
Community	12.9k	4.0k	1297	38.70
Organisation	24.9k	6.7k	382	31.74

Table 2: Text Length Distribution for 'Individual', 'Organisation', and 'Community' Classes.

In the preprocessing step, the initial transformation involves extracting text from images using the Google Vision API. Subsequently, various techniques are applied to clean and refine the text data before inputting it into the model. The process begins by eliminating any HTML tags present in the text, followed by the normalization of accent characters to their ASCII equivalents. Punctuation marks and special characters are eliminated, and the entire text is converted to lowercase for uniformity. The text is subsequently tokenized into individual words and any unnecessary spaces are stripped. This preprocessing step aims to prepare the text data in a standardised and meaningful for-

mat, enhancing the model's performance. For subtask 1, Table 3 presents a comparison between the unprocessed text in column 1 and the processed text in column 2, along with their respective labels in column 3. For subtask 2, a similar comparison is shown in Table 4, where column 1 contains the unprocessed text, column 2 displays the processed text, and column 3 indicates the corresponding labels. The preprocessing step plays a vital role in optimising the input data for the hate speech detection models, ensuring that the models can effectively capture the relevant patterns and features from the text-embedded images.

Extracted Text From Images	Preprocessed text	Label
BREAKING NEWS: PRESIDENT ZELENSKYY TRIPPED AND FELL HERE THIS MORNING imglip.com	breaking news president zelenskyy tripped and fell here this morning imglipcom	Hate Speech
PROTESTORS AROUND THE WORLD RALLY IN SUPPORT OF UKRAINE STORYFUL/AP	protestors around the world rally in support of ukraine storyfulap	No Hate Speech

Table 3: Example of text extracted from CrisisHateMM dataset for sub-task 1

Extracted Text From Images	Preprocessed text	Label
HEY JOE, RUSSIA HAS INVADED UKRAINE WITH HEAVY ARSENAL! WHAT YOU GONNA DO?	hey joe russia has invaded ukraine with heavy arsenal what you gonna do im holding a climate denialism roundtable imgflipcom	Individual
58 13 CCM All saver 'WISHING THEM DEATH' Russian NHL players face horrible anger	58 13 ccm all saver wishing them death russian nhl players face horrible anger	Community
THE WEST HAS GIVEN THE BEST SONG AWARD TO NATO NATO imgflip.com	the west has given the best song award to nato nato imgflipcom	organisation

Table 4: Example of text extracted from CrisisHateMM dataset for sub-task 2

3.2 Model training and evaluation

At the classification stage, four deep learning models, namely Xlm-RoBERTa-base, BiLSTM, XLNet base cased, and ALBERT—are used to categorise posts containing hate speech based on the pre-processed text. The ability of these models to extract complex patterns and contextual information from the textual data is a commonality shared by them. To do this, they employ advanced language representation techniques and attention mechanisms. A thorough evaluation of their capacity to correctly identify hate speech within the text-embedded images is provided by the use of critical metrics to measure each participant's performance, including accuracy, F1 score, precision, and recall. To ensure their dependability and suitability for the crucial task of identifying hate speech, these models go through extensive testing and analysis against the established metrics.

3.2.1 Bidirectional LSTM

Bidirectional Long Short-Term Memory (BiLSTM) is based on LSTM architecture that processes input data in both forward and backward directions. The hidden state in typical LSTM is updated based on previous information in the input sequence (i.e., from left to right). Bidirectional LSTM, on the other hand, processes the input sequence in two passes: one from left to right (ahead direction) and one from right to left (reverse way). The capacity of Bidirectional LSTM to capture information from both past and future contexts is a critical advantage for comprehending the context and dependencies in a sequence. Because the model is bidirectional, it can capture long-term dependencies and context that traditional unidirectional LSTMs may miss.

3.2.2 Xlnet base cased

'xlnet-base-cased' is a variation of the XLNet language model that is part of Google Research's Transformer-based family. It enhances the BERT model by using a permutation-based training strategy to overcome some shortcomings. It is suited for a variety of natural language processing tasks such as text categorization, sentiment analysis, language synthesis, and question answering after being trained on a large corpus. Because it is 'cased,' it keeps casing information in input text, which is useful for some jobs. Researchers and developers can fine-tune the 'xlnet-base-cased' model for specific tasks by using its rich language representation capabilities, which capture complex linguistic

patterns and context for a wide range of natural language processing applications.

3.2.3 ALBERT-base-v2

ALBERT-base-v2 is a language model variant of ALBERT (A Lite BERT), aiming to be a more efficient and parameter-reduced version of BERT. The suffix "base" denotes that it is smaller than larger variants such as 'ALBERT-large' or 'ALBERT-xlarge'. ALBERT achieves efficiency by utilising parameter-sharing techniques such as factorised embeddings and cross-layer parameter sharing, which results in quicker training times without sacrificing performance. It is trained using masked language modelling (MLM) on a huge corpus and may be fine-tuned for various NLP tasks. ALBERT has grown in prominence due to its competitive performance, particularly in resource-constrained circumstances, and is now a common choice in NLP applications.

3.2.4 Xlm-RoBERTa-base

Xlm-RoBERTa-base' is a variant of the XLM-R (Cross-lingual Language Model - Roberta) language model. It is a multilingual version of the RoBERTa model, based on the transformer architecture, designed for cross-lingual language understanding. The model is trained on a large corpus of text data from multiple languages using masked language modeling (MLM) and translation language modeling (TLM) objectives. This allows it to effectively process and understand text from various languages. The "cased" in the name indicates that it retains case information during training and inference, treating uppercase and lowercase characters as distinct tokens. XLM-Roberta-base is widely used in multilingual NLP tasks (Conneau et al., 2020), transferring knowledge across languages and performing well on tasks involving different languages.

4 Results and discussion

Classification results of Sub-task 1 are reported in Table 5 for all four classifiers. XLM-Roberta-base has outperformed all other classifiers in terms of all four metrics. XLM-Roberta-base has been trained on large corpses of text data hence it can understand the context of text more effectively as compared to the other classical NLP models. From the values of the F1-score, it can be concluded that the models have learned the context of both classes and performed well to identify each class.

	BiLSTM	ALBERT	XLnet	XLM-Roberta
Acc.	68.62	81.71	82.84	84.65
F1	68.62	81.56	82.78	84.62
Recall	69.00	81.60	83.03	85.07
Prec.	68.86	81.53	82.74	84.76

Table 5: Model Performance for Hate Speech Detection (sub-task 1).

The classification results for sub-task 2 are summarised in Table-6, utilising the same four classifiers as in sub-task 1. Once again, XLM-Roberta-base stands out as the best performer, surpassing the other classifiers in all four evaluation metrics. This exceptional performance can be credited to its extensive training on text data, which enables it to comprehend the context of textual content more effectively than traditional NLP models. The Precision values further validate that the models have successfully achieved accurate target classification for individual, community, and organisational categories. These results reaffirm the significance of XLM-Roberta-base in hate speech detection and target classification within text-embedded images, underlining its potential for advancing research in this field. XLM-Roberta-base's superior performance in hate speech detection for both sub-tasks, outperforming other classifiers. Its extensive pre-training on vast text corpora, bidirectional context comprehension, large capacity, multilingual proficiency, and fine-tuning on CrisisHateMM dataset contribute to its exceptional understanding of hate speech content. Ethical considerations and challenges in detecting hate speech were acknowledged. The CrisisHateMM dataset's value for research, providing insights into hate speech complexities, was emphasised. Leveraging advanced NLP models like XLM-Roberta-base holds significant potential for effective hate speech detection and content moderation, fostering a safer online environment.

	BiLSTM	Albert	XLnet	XLM-Roberta
Acc.	56.19	67.35	66.52	72.23
F1	54.84	65.35	62.32	69.73
Recall	58.99	65.35	61.56	68.94
Prec.	59.99	65.36	64.47	71.01

Table 6: Performance Comparison of NLP Models for Target Identification (sub-task 2).

Furthermore, our achievement of the 3rd rank in both sub-tasks using solely textual models, as evident in Table 7 and Table 8, not only underscores the efficiency of the XLM-Roberta-base model but

Team Name	Recall	Precision	F1	Accuracy
ARC-NLP	85.67	85.63	85.65	85.78
bayesiano98	85.61	85.28	85.28	85.33
IIC Team	**85.08**	**84.76**	**84.63**	**84.65**
DeepBlueAI	83.56	83.35	83.42	83.52

Table 7: In Subtask A: Hate Speech Detection leaderboard, our team, IIC Team, ranks third based on the F1 score, demonstrating competitive performance across all classification metrics.

Team Name	Recall	Precision	F1	Accuracy
ARC-NLP	76.36	76.37	76.34	79.34
bayesiano98	73.30	75.54	74.10	77.27
IIC Team	**68.94**	**71.05**	**69.73**	**72.31**
Sarika22	67.77	68.41	68.05	71.49

Table 8: In Subtask B: Target Detection, our team, IIC Team, secured the third position on the leaderboard, leading in all classification metrics based on the F1 score.

also highlights the prudent management of computational resources. This approach aptly aligns with resource-conscious strategies, demonstrating a commitment to optimizing performance while maintaining a responsible balance.

5 Conclusion

This research focuses on text-embedded images used in social media to express opinions and emotions, which unfortunately also serve as platforms for spreading hate speech, propaganda, and extremist ideologies. Notably, during the Russia-Ukraine war, both sides extensively utilised text-embedded images for propaganda and hate speech dissemination. The growing abundance of offensive content on social media poses challenges in effectively detecting and moderating such material. To address this issue, we utilise the CrisisHateMM dataset, an innovative multimodal dataset containing over 4,700 text-embedded images from the Russia-Ukraine conflict. The dataset is meticulously annotated for hate and non-hate speech, further categorising hate speech into directed and undirected forms and providing annotations for individual, community, and organisational targets. Our research involves two subtasks: Sub-task 1 focuses on hate speech detection in text-embedded images, while Sub-task 2 aims to identify the targets of hate speech. To achieve accurate results, we employ advanced feature extraction techniques and utilise deep learning models for both subtasks, yielding promising outcomes. In Sub-task 1, our textual model, XLM-Roberta-base, demonstrated superior performance, achieving the highest accuracy on test(unseen) data with a recall of 85.08%, precision of 84.76%, F1 score of 84.63%, and accuracy of 84.65%. Additionally, in Sub-task 2, the Xlm-Roberta-base model outperformed other approaches, achieving a recall of 68.94%, precision of 71.05%, F1 score of 69.73%, and accuracy of 72.31% These results highlight the effectiveness of our approach in hate speech detection and target identification in text-embedded images during the Russia-Ukraine conflict. Exploring multimodal approaches, accessing larger and more diverse datasets, fine-tuning strategies, addressing biases, integrating automated systems with social media platforms, extending detection to multiple languages, and improving interpretability and contextual understanding are all part of the future of hate speech detection and content moderation. Exploring zero-shot and few-shot learning methodologies, as well as addressing ethical concerns, are also essential. In summary, future research promises effective and responsible ways for combating hate speech on social media through the use of AI developments.

References

Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Information*, 13(6):273.

James Banks. 2010. Regulating hate speech online. *International Review of Law, Computers & Technology*, 24(3):233–239.

David M Beskow, Sumeet Kumar, and Kathleen M Carley. 2020. The evolution of political memes: Detecting and characterizing internet memes with multi-modal deep learning. *Information Processing & Management*, 57(2):102170.

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.

Sergio Andrés Castaño-Pulgarín, Natalia Suárez-Betancur, Luz Magnolia Tilano Vega, and Harvey Mauricio Herrera López. 2021. Internet, social media and online hate speech. systematic review. *Aggression and Violent Behavior*, 58:101608.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. 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.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Liam Hebert, Gaurav Sahu, Nanda Kishore Sreenivas, Lukasz Golab, and Robin Cohen. 2023. Multi-modal discussion transformer: Integrating text, images and graph transformers to detect hate speech on social media. *arXiv preprint arXiv:2307.09312*.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

B Mathew, R Dutt, P Goyal, and A Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science*. Anais.

Ahmed Cherif Mazari, Nesrine Boudoukhani, and Abdelhamid Djeffal. 2023. Bert-based ensemble learning for multi-aspect hate speech detection. *Cluster Computing*, pages 1–15.

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.

Gretel Liz De la Pena Sarracén, Reynaldo Gil Pons, Carlos Enrique Muniz Cuza, and Paolo Rosso. 2018. Hate speech detection using attention-based lstm. *EVALITA evaluation of NLP and speech tools for Italian*, 12:235.

Aneri Rana and Sonali Jha. 2022. Emotion based hate speech detection using multimodal learning. *arXiv preprint arXiv:2202.06218*.

ES Smitha, Selvaraju Sendhilkumar, and GS Mahalaksmi. 2018. Meme classification using textual and visual features. In *Computational Vision and Bio Inspired Computing*, pages 1015–1031. 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, case 2023. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*.

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.