A Multi-Modal Dataset for Hate Speech Detection on Social Media: Case-study of Russia-Ukraine Conflict

Surendrabikram Thapa[1], Aditya Shah[1], Farhan Ahmad Jafri[2], Usman Naseem[3], Imran Razzak[4]
[1]Department of Computer Science, Virginia Tech, USA
[2]Department of Computer Science, Jamia Millia Islamia, India
[3]School of Computer Science, The University of Sydney, Australia
[4]School of Computer Science and Engineering, University of New South Wales, Australia
{sbt,aditya31}@vt.edu, farhanjafri88888@gmail.com
usman.naseem@sydney.edu.au, Imran.razzak@unsw.edu.au

Abstract

Hate speech consists of types of content (e.g. text, audio, image) that express derogatory sentiments and hate against certain people or groups of individuals. The internet, particularly social media and microblogging sites, have become an increasingly popular platform for expressing ideas and opinions. Hate speech is prevalent in both offline and online media. A substantial proportion of this kind of content is presented in different modalities (e.g. text, image, video). Taking into account that hate speech spreads quickly during political events, we present a novel multimodal dataset composed of 5680 text-image pairs of tweets data related to the Russia-Ukraine war and annotated with a binary class: "hate" or "no-hate" The baseline results show that multimodal resources are relevant to leverage the hateful information from different types of data. The baselines and dataset provided in this paper may boost researchers in direction of multimodal hate speech, mainly during serious conflicts such as war contexts.

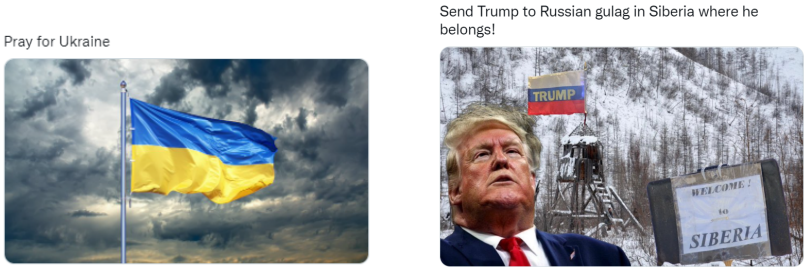
1 Introduction

The internet has become an increasingly popular communication medium to express the views of people. People mostly express their opinions on various topics using social media, microblogging platforms, blogs, etc. With great internet penetration even in the rural parts of the world and ease of access to information in real-time, people mostly rely on social media platforms (Naseem et al., 2021). At times of political events and tension in any region, the users of such platforms become more active than usual and post their thoughts and updates regarding the issues. During the expression of such opinions and ideas, there can be mixed emotions. Some opinions lean towards supporting the people on the ground who are suffering in such political events whereas some opinions are about blaming each other, name-calling, exaggeration of information, etc (Dimitrov et al., 2021). In political situations pertaining to invasion, the situation becomes even worse. Social media sometimes get polarized into the ones supporting the invasion and the ones opposing the invasion. During such polarization, a lot of content can be found which uses extreme language, falsifies the information, and spreads hate. Such content when directed towards certain people or groups of individuals (race, gender, nationality) with the intent to show anger and hate is called hate speech (Parihar et al., 2021). While the legal definitions of hate speech vary from territory to territory, hate speech on the internet sphere is taken as hateful content on the internet that is directed toward certain individuals or groups of individuals. The Cambridge Dictionary defines hate speech as "public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex, or sexual orientation" (Miller and Brown, 2013).

On February 24, 2022, Russia started a full-scale invasion of Ukraine by land, sea, and air (Berninger et al., 2022). The world was again polarized into two, with one supporting the Russian invasion and the other opposing it. Many countries condemned the war, and sanctions were eventually imposed on Russia. With the development of these events, social media started getting active. People started to express their opinions related to the humanitarian crisis and economic crisis that was caused due to the invasion. Amid the healthy and respectful discourse and discussions, there was some hateful content targeted at various people (Figure 1).

Hate speech can bring serious consequences to society. Microblogging platforms and social media platforms put a lot of effort into managing the hateful content on their platforms. Mostly, the platforms use human mediators for the mediation of posts related to hate speech. Despite being an efficient method for regulating hate speech, it is not always possible for human mediators to flag

Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE), pages 1 - 6



(a) Tweet with No Hate (b) Tweet with Hate Speech

Figure 1: Examples of tweets with hate and no hate speech during Russia-Ukraine conflict

the posts provided that the volume of the hateful content becomes extremely high in situations of political events like an invasion. Thus, there has always been a need for an automated system to identify the contents related to hate speech. Our ***contributions*** can be summarized as follows:

- We construct and release new multi-modal data for identifying hate speech tasks on social media, consisting of 5,680 tweets (image-text pairs) labeled across binary labels.
- Our experimental analysis shows that both modalities (text and images) are important for the task.
- We experiment with several state-of-the-art textual, visual, and multi-modal models, which further confirm the importance of both modalities and the need for further research.

2 Related Works

Despite hate speech detection being a hard task, much research is being done to address hate speech on the internet. With advancements in the field of deep learning, there is a multitude of problems that are being solved by deep learning (Adhikari et al., 2022). Hate speech is one of the tasks that is being explored using deep learning techniques. Most of the research on hateful content is focused on leveraging the information from the textual content. Del Vigna et al. (2017) curated a dataset of 17,567 comments from Facebook posts and annotated for strong hate, weak hate, and no hate categories. The proposed long-short term memory (LSTM) (Hochreiter and Schmidhuber, 1997) and SVM models performed with an accuracy of 72.95% and 75.23% for hate and non-hate categories. Similarly, the accuracy of 64.61% and 60.50% were reported across all three categories. Similarly, Gambäck and Sikdar (2017) had proposed multiple CNN architectures in order to classify hate speech spanning across multiple classes, viz. racism, sexism, both (racism and sexism), and non-hate speech. The architecture with Word2Vec embeddings was able to achieve an F1-score of 0.7829. Calderón et al. (2020) did a slightly different task of hate speech classification by curating a dataset (1977 tweets) of the hate speech directed towards the immigrants in Spain and performing the task of topic modeling and meticulously studying linguistic cues of hate speech.

Apart from these, some research has been done on multi-modal hate speech detection. For instance, Shang et al. (2021) proposed Analogy-aware Offensive Meme Detection (AOMD) that was able to learn the implicit analogy from the multi-modal contents of the meme and detect the offensive analogy. The model that used ResNet50 (He et al., 2016) and Glove-based LSTM was able to achieve the accuracy of 69% and 72% for Gab and Reddit datasets. Similarly, Zhou et al. (2021) proposed a method that integrates the image captioning process into the memes detection process. The approach enhanced the cross-modality relationship and helped achieve AUROC as high as 78.86. For their study, they used the famous dataset from Hateful Memes Challenge (Kiela et al., 2020). Similarly, Dimitrov et al. (2021) presented a method to identify propaganda techniques in memes by leveraging the multi-modal information and classifying them into 22 propaganda techniques.

In recent days, the research relating to multi-modal information has been growing (Sharma et al., 2022). Most microblogging sites allow users to post in various modalities like text, images, videos, etc. which add a dimension of research in addressing all the modalities. One modality often provides supplementary information to another modality which makes multimodal models more robust.

3 Datasets

3.1 Data Collection

The Russian invasion of Ukraine started on 22 February 2022. We started to crawl tweets from

Label	Annotation Instructions
Hate	A post (text or image or both) contains a hateful content such as personal attack, homophobic abuse, racial abuse, or attack on minority
No Hate	A post(text or image or both) reports the events or others' opinions objectively and contains no offensive or hateful content.

Table 1: Annotation instructions given to annotators.

22 February 2022 to 28 March 2022. Twitter API[1] was used to collect the tweets from the given time frame. We collected the tweets with certain list of keywords namely *ukraine*, *putin*, *russia*, *zelensky*, *kyiv*, *kiev*, *kremlin*, *ukrainian*, *nato*, *russian*, *soviet*, *moscow*, *kharkiv*, and *donbas*. The tweets for keywords *kharkiv*, and *donbas* were collected from 1 March 2022 whereas for all other keywords, tweets were collected starting from 22 February 2022. The tweets revolving around the Russia-Ukraine crisis had the above-mentioned keywords very frequently. Hence, the mentioned keywords were selected for our study. For filtering the tweets, we took the tweets which had media and were in the English language. We discarded the tweets which had media as videos or animations. Our dataset contains 5,680 labeled tweets that had image and text pairs with annotations.

3.2 Annotation

This subsection explains the annotation schema that we followed to label the dataset.

Instructions: The annotation of the data was done to label tweets into binary classes. The two categories, i.e., hate speech and no hate speech, were defined. Annotators were provided with the instructions, following which they assigned the labels to the tweets. If the annotators were not sure about the labels for any tweet, it was labeled as 'Non-Informative,' and such tweets were later dropped. Annotators were provided with posts that had tweets containing both image and text pairs. The images were named as the tweet ID in which they were present. The annotators thus looked into the image and text pairs for performing the annotation. Annotation instructions given to annotators are presented in Table 1. For a tweet to be labeled as hate speech, it needs to have at least one component that represents hate.

Annotations: There was a team of four male and female annotators with good fluency in the English language. All annotators had varying qualifications running from undergraduate to MS and Ph.D. degrees, including the highly experienced researchers in NLP research involving the data collection and establishment of benchmarks. This helped to frame clear instructions and ensure the quality of annotations. In the literature, it has been discussed that having a diverse range of annotators is useful to mitigate bias (Vargas et al., 2022). The annotators were volunteers and did not receive any remunerations. Since labeling tweets involving both text and image is challenging, we made the annotations go through three phases. In the first phase, we run a pilot annotation for 50 tweets to ensure that everyone understood the instructions. Each of the four members annotated the tweets. The instructions were revised to clarify that they addressed all the confusion that annotators had. In the second phase, all four annotators were made to annotate 200 tweets. The purpose of the second phase was to make sure that the instructions revised after the first stage were clear enough. In the third stage, a group discussion was done regarding the conflicts in annotation (Table 3). The instructions became apparent, and the annotators annotated all of the datasets. For example, Figure 1a shows that the text expresses solidarity with Ukraine. The image, which is the flag of Ukraine, also does not show any hate. Thus, the tweet is labeled as No Hate. Similarly, 1b shows the tweet in which the text shows hate towards the former president of the USA, Donald Trump. He does not belong to Siberia. The tweet text tries to demean Donald Trump by saying that he belongs to Siberia and he should be sent there. The image is also edited. It is demeaning and shows hate on multiple levels toward Donald Trump. Thus, this is labeled as hate speech.

Labels	No. of Tweets	Avg. char/ tweet	Avg word/ tweet
Hate	746	60.88	9.68
No Hate	4934	64.48	10.03

Table 2: Dataset statistics.

Dataset Statistics and Analysis: Our new multimodal dataset included 5680 tweets, with 746 (13.13%) tweets being labeled as 'hate speech' label whereas 4934 (86.87%) tweets are labeled as 'no hate' label (Table 2). The dataset statistics represent a true distribution in a real-world scenario

[1] https://developer.twitter.com/en/docs/twitter-api

Phase	Annotators	Kappa (κ)
Pilot Annotation	α_1 and α_2	0.57
	α_1 and α_3	0.50
	α_1 and α_4	0.62
	α_2 and α_3	0.53
	α_2 and α_4	0.63
	α_3 and α_4	0.51
Final Annotation	α_1 and α_2	0.87
	α_1 and α_3	0.90
	α_1 and α_4	0.89
	α_2 and α_3	0.89
	α_2 and α_4	0.88
	α_3 and α_4	0.90

Table 3: Cohen's Kappa (κ) for annotation during different Phases by four annotators

where many posts are neutral, and only some are related to hate speech.

4 Experimental Results

4.1 Baselines

We used various state-of-the-art unimodal and multimodal-based state-of-the-art methods to establish baselines. Below, we discuss each in detail.

4.1.1 Unimodal Models

For single modality-based models, we used the following unimodal methods:

- Unimodal-Text Only: For textual models, we used long-short term memory (LSTM) (Hochreiter and Schmidhuber, 1997), Bidirectional Encoder Representations (BERT) (Devlin et al., 2018) and optimized variant of BERT, i.e., RoBERTa (Liu et al., 2019).

- Unimodal-Image Only: For the image-based unimodal baseline methods, we used 3 pretrained convolutional networks based methods i.e., VGG-19 (Simonyan and Zisserman, 2014), ResNet (He et al., 2016) and DenseNet (Huang et al., 2017).

4.1.2 Multimodal Models

We used 3 multimodal models that have been widely used in previous similar studies. (1) We used (ResNet+BERT), where we pre-trained ResNet and BERT to train text and image and then fused the representations through the linear layer, (2) We also used VisualBERT (Li et al., 2019), a simple and flexible framework for modeling a broad range of vision-and-language tasks and (3) Besides, we have also used the current state-of-the-art model Contrastive Language-Image Pre-Training (CLIP) (Radford et al., 2021).

Modality	Model	Precision	Recall	F1-score
Textual	LSTM	0.74	0.86	0.79
	BERT	0.75	0.86	0.80
	RoBERTa	0.78	0.88	0.83
Visual	VGG-19	0.79	0.70	0.74
	ResNet	0.80	0.74	0.77
	DenseNet	0.82	0.72	0.77
Multimodal	ResNet+BERT	0.84	0.86	0.85
	VisualBERT	0.85	0.88	0.86
	CLIP	0.88	0.90	0.89

Table 4: Performance of different unimodal and multimodal algorithms on our dataset.

4.2 Experimental Settings

We used grid-search optimization to derive the optimal parameters of each baseline and used precision, recall, and F1-score as evaluation metrics.

4.3 Results

Table 4 show the results for the classification of hate and non-hate speech. We experimented with both unimodal and multimodal models. When only the text modality was used, the RoBERTa model performed the best with an F1-score of 0.83. Similarly, for the visual unimodal model, DenseNet and ResNet had a nearly equal performance with an F1-score of 0.77. Further, we can see that both multimodal models had better results than unimodal textual and visual models. The performance for the CLIP model is as high as 0.89 (F1-score). Based on our experiment, we observed that multi-modal models plays important role in detecting hateful content in comparison to uni-models.

5 Conclusion and Future Work

This paper presents a new multi-modal dataset for identifying hateful content on social media, consisting of 5,680 text-image pairs collected from Twitter, labeled across two labels. Experimental analysis of the presented dataset has shown that understanding both modalities is essential for detecting these techniques. It is confirmed in our experiments with several state-of-the-art multi-modal models. In future work, we plan to extend the dataset in size. We further plan to develop new multi-modal models tailored explicitly to hate-speech detection, aiming for a deeper understanding of the text and image relation. It would also be interesting to perform experiments in a direction that explores what social entities the given hate speech tweet targets.
Reproducibility: The dataset and resources for this work are available at our GitHub repository[2].

[2]https://github.com/therealthapa/emnlp-case2022

Ethical Considerations: The dataset does not contain direct identifiers. It contains tweet IDs. Tweet IDs can be used to retrieve the tweets. The tweet becomes unavailable if the user deletes the tweet. This gives the original author of the tweet full control over their content. All the tweets presented in the examples have been anonymized and obfuscated for user privacy and to avoid misuse. Thus, no ethical approval is required. The annotation is very subjective and hence we can expect some bias in the annotation. To address these issues, examples from various users and groups are collected, along with clear instructions for annotation. Due to excellent inter-annotator agreement (κ score), we are confident that annotation instructions are mostly valid.

Intended Use: We release our dataset in order to accelerate research into identifying hate speech at times of war on social media. We expect the dataset to be a valuable resource when used appropriately.

References

Surabhi Adhikari, Surendrabikram Thapa, Usman Naseem, Priyanka Singh, Huan Huo, Gnana Bharathy, and Mukesh Prasad. 2022. Exploiting linguistic information from nepali transcripts for early detection of alzheimer's disease using natural language processing and machine learning techniques. *International Journal of Human-Computer Studies*, 160:102761.

Marc Berninger, Florian Kiesel, and Sascha Kolaric. 2022. Should i stay or should i go? stock market reactions to companies' decisions in the wake of the russia-ukraine conflict. *Stock market reactions to companies' decisions in the wake of the Russia-Ukraine conflict (April 20, 2022)*.

Carlos Arcila Calderón, Gonzalo de la Vega, and David Blanco Herrero. 2020. Topic modeling and characterization of hate speech against immigrants on twitter around the emergence of a far-right party in spain. *Social Sciences*, 9(11):188.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Detecting propaganda techniques in memes. *arXiv preprint arXiv:2109.08013*.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33:2611–2624.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

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*.

Jim E Miller and E Keith Brown. 2013. *The Cambridge dictionary of linguistics*. Cambridge University Press.

Usman Naseem, Imran Razzak, Matloob Khushi, Peter W Eklund, and Jinman Kim. 2021. Covidsenti: A large-scale benchmark twitter data set for covid-19 sentiment analysis. *IEEE Transactions on Computational Social Systems*, 8(4):1003–1015.

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. *CoRR*, abs/2103.00020.

Lanyu Shang, Yang Zhang, Yuheng Zha, Yingxi Chen, Christina Youn, and Dong Wang. 2021. Aomd: An analogy-aware approach to offensive meme detection on social media. *Information Processing & Management*, 58(5):102664.

Shivam Sharma, Firoj Alam, Md 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. *arXiv preprint arXiv:2205.04274*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. 2022. Hatebr: A large expert annotated corpus of brazilian instagram comments for offensive language and hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183.

Yi Zhou, Zhenhao Chen, and Huiyuan Yang. 2021. Multimodal learning for hateful memes detection. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE.