CrisisHateMM: Multimodal Analysis of Directed and Undirected Hate Speech in Text-Embedded Images from Russia-Ukraine Conflict

Aashish Bhandari[1§], Siddhant B. Shah[1§], Surendrabikram Thapa[2§], Usman Naseem[3], Mehwish Nasim[4, 5]
[1]Department of CSE, Delhi Technological University, India
[2]Department of Computer Science, Virginia Tech, USA
[3]School of Computer Science, The University of Sydney, Australia
[4]School of Physics, Mathematics and Computing, The University of Western Australia, Australia
[5]College of Science and Engineering, Flinders University, Australia

Abstract

Text-embedded images are frequently used on social media to convey opinions and emotions, but they can also be a medium for disseminating hate speech, propaganda, and extremist ideologies. During the Russia-Ukraine war, both sides used text-embedded images extensively to spread propaganda and hate speech. To aid in moderating such content, this paper introduces CrisisHateMM, a novel multimodal dataset of over 4,700 text-embedded images from the Russia-Ukraine conflict, annotated for hate and non-hate speech. The hate speech is annotated for directed and undirected hate speech, with directed hate speech further annotated for individual, community, and organizational targets. We benchmark the dataset using unimodal and multimodal algorithms, providing insights into the effectiveness of different approaches for detecting hate speech in text-embedded images. Our results show that multimodal approaches outperform unimodal approaches in detecting hate speech, highlighting the importance of combining visual and textual features. This work provides a valuable resource for researchers and practitioners in automated content moderation and social media analysis. The CrisisHateMM dataset and codes are made publicly available at https://github.com/aabhandari/CrisisHateMM.

1. Introduction

The widespread usage of social media has resulted in a significant increase in multimodal data [1]. This data type includes various forms of content, such as text, images, and video. The volume of such data has grown substantially, with a 3% increase in social media users from 2022 to 2023, bringing the total number of users to 4.76 billion worldwide [2]. However, along with the growth in usage, the amount of offensive and hateful data has also increased. The rise of offensive and hateful content on social media poses several challenges for detecting and moderating such content [3]. Traditional approaches to content moderation, such as manual review and filtering, are no longer effective due to the sheer volume of data. Therefore, automated approaches are needed to detect and remove offensive and hateful content. Apart from making content moderation easier, automated hate speech moderation can also reprieve social media moderators from over-exposure to hateful content, which can be psychologically damaging [4]. However, one significant challenge is the development of algorithms that can recognize different forms of multimodal data. Most current algorithms are designed to analyze text-based content and struggle to identify and moderate images and video [5]. To address this issue, researchers need to develop new algorithms to analyze multiple data forms, such as image recognition and audio analysis. However, the limited data available to train algorithms has impeded the ability to detect and moderate content automatically, especially during political events like invasions.

The Russia-Ukraine conflict that started on February 24, 2022, triggered a wave of social media activity [6, 7]. The conflict saw polarized opinions, with one side supporting the Russian invasion and the other opposing it. Social media provided a platform for people to express their views on the conflict, but it also led to the spread of hateful content. Text-embedded images were widely used to disseminate hate speech on social media. Russian state media and pro-Russian separatists used such images to portray Ukrainians as fascists and Nazis, while Ukrainian activists and supporters used them to highlight Russian aggression

[§]The authors contributed equally and are joint first authors. The ordering of authors is alphabetical.
WARNING: SOME EXAMPLES PRESENTED IN THIS PAPER CONTAIN HATEFUL CONTENT AND MIGHT BE OFFENSIVE.

(a) No Hate (b) Undirected Hate (c) Target: Individual (d) Target: Organization (e) Target: Community

Figure 1. Examples of text-embedded images labeled for hate speech and sub-classes from CrisisHateMM dataset. Text-embedded images for directed hate speech were further annotated for target classes, as shown in the figure.

and human rights violations, as shown in Figure 1. Such hate speech can exacerbate ongoing tensions and potentially incite more violence. Therefore, it is essential to identify instances of hate speech, especially in determining whether it is directed towards a particular target, to address the issues that arise from such content. Detecting targets would help protect vulnerable groups and plan for specific interventions. However, despite the severity of this issue, there has been limited research on hate speech detection during political events such as the Russia-Ukraine war.

Bridging this gap, we investigate hate speech in text-embedded images in social media and the internet. We annotate a unique dataset related to the Russia-Ukraine crisis to address three main tasks: (i) **Task A:** Detecting whether a given text-embedded image is hateful or not. (ii) **Task B:** Detecting whether the hate speech is directed or undirected. (iii) **Task C:** Detecting the targets of directed hate speech in given text-embedded images. Our main contributions are:

- We create and release a dataset of 4,723 text-embedded images manually annotated to identify hate speech, the direction of hate speech, and the targets of directed hate speech.
- We do a preliminary analysis of the data and benchmark the dataset with various textual, visual, and multimodal algorithms.
- Our experiments show that multiple modalities are important to better understand hate speech in text-embedded images.

2. Related Work

In recent years, the identification of hate speech on social media has become an important research topic in the field of computational linguistics [15]. However, one of the main challenges in this area has been the lack of relevant data, which has hindered the development of effective methods for detecting hate speech. To tackle this problem, researchers have been curating novel datasets with the purpose of aiding the identification of hate speech on social media. Fortuna et al. [8] proposed a dataset of 5,668 tweets annotated into 81 categories of hate speech in the Portuguese language. Similarly, Pereira-Kohatsu et al. [9] curated a dataset of 6,000 Spanish tweets on hate speech along with an unlabeled corpus of 2 million tweets.

Political events may have a significant impact on society as they often have the ability to sway public opinion on a large scale. Thus, it is important to curate datasets that have a relevant political context. Kumar et al. [10] proposed TweetBLM, a dataset related to the Black Lives Matter movement, which was manually annotated for hate speech. Similarly, Grimminger et al. [11] introduced a dataset consisting of 3,000 tweets related to the 2020 US elections, categorized them according to their political stance toward a candidate, and further classified those tweets as offensive and non-offensive.

While the vast majority of research in hate speech detection remains limited to unimodal methods, the research in text-embedded images is equally important because of the ease of sharing such content. Most research accounts for either visual or textual information for detecting hate speech in text-embedded images. Leveraging multimodal information, typically the combination of textual and visual information, has proven to robustly detect hate speech in social media for multimodal content [16]. Liu et al. [17] introduced Figmemes, a multimodal dataset consisting of 5,141 politically-opinionated memes annotated according to the figurative language used in them. Similarly, Gasparini et al. [12] collected multimodal data of 800 memes categorized into misogynistic, ironic, and aggressive content.

The Russia-Ukraine conflict emphasizes the role social media plays in modern-day warfare. Alongside the conflict in the physical environment, there was an active conflict in the information environment. This prompted numerous research efforts to monitor user trends and moderate hate speech. Smart et al. [18] collected a dataset of over $5Mn$ tweets, to quantify how bots were influencing people in the online conversation around the Russia-Ukraine conflict. Hasan et al. [19] collected a dataset containing 10,861 Bengali comments regarding the Russia-Ukraine crisis posted on YouTube news channels and annotated them into three categories: 'Pro-Ukraine', 'Pro-

Table 1. Summary of datasets used in the literature related to hate speech detection. CrisisHateMM (our dataset) is multimodal and has different sub-classes. It is the first dataset annotating text-embedded images that has the context of the Russia-Ukraine crisis.

Work	Data Source	Multimodal	Sub-classes	Size	Context
Fortuna et al. [8]	Twitter	✗	✓	5668	✗
Pereira-Kohatsu et al. [9]	Twitter	✗	✗	6000	✗
Kumar et al. [10]	Twitter	✗	✗	9165	BLM movement
Grimminger et al. [11]	Twitter	✗	✓	3,000	2020 US Elections
Gasparini et al. [12]	Facebook, Twitter, Instagram, Reddit	✓	✗	800	Online Misogyny
Kiela et al. [13]	Self Generated	✓	✗	10,000	✗
Thapa et al. [14]	Twitter	✓	✗	5,680	Russo-Ukraine War
CrisisHateMM (Ours)	Twitter, Facebook, Reddit	✓	✓	4,723	Russo-Ukraine War

Russia', and 'Neutral'. Similarly, Toraman et al. [20] curated a dataset containing 5,284 English and 5,064 Turkish tweets pertaining to recent topics such as the Russia-Ukraine war, COVID-19, and Refugees, where the misinformation propagated by the tweets was analyzed. Thapa et al. [14] curated a multimodal dataset containing 5,680 image-caption text pairs obtained from tweets regarding the Russia-Ukraine conflict. They categorized their data into two classes: Hate and Non-Hate, and their experiments reflected the superiority of multimodal algorithms over unimodal visual and text methods. However, such image-caption pairs may have a lower degree of congruence between visual and textual data than text-embedded images, which may cause processing systems to capture inaccurate contextual information. Furthermore, existing datasets are restricted solely to one social media platform, and thus they may only represent the opinions of a small percentage of the population. Moreover, existing works suffer from limitations such as a lack of subclassing during annotation, reliance on unimodal data, and using single-platform data. To fill this void, we annotate text-embedded images from various platforms for different sub-categories. We hope that our comprehensively annotated dataset acts as a stepping stone toward robust and cross-platform content moderation on social media. Table 1 provides a detailed comparison of the hate speech datasets used in the literature. Unlike other existing datasets, which are either unimodal or lack sub-classes and context, our dataset is multimodal with hierarchical annotation of sub-classes.

3. Dataset

Text-embedded images refer to images that contain textual information within the image itself. This text can be used to provide additional context, or explanation, or to convey a specific message related to the image. Examples of text-embedded images include infographics, social media posts, memes, news snippets, and posters. For our dataset, we curated text-embedded images starting from February 24, 2022, the day the president of Russia, Vladimir Putin, announced the initiation of a special military operation in Ukraine, to March 3, 2023. This section provides the specifications of the data collection process along with the annotation guidelines that were used to annotate the text-embedded images. A schematic overview of the data collection, along with the hierarchical annotation process, is shown in Figure 2.

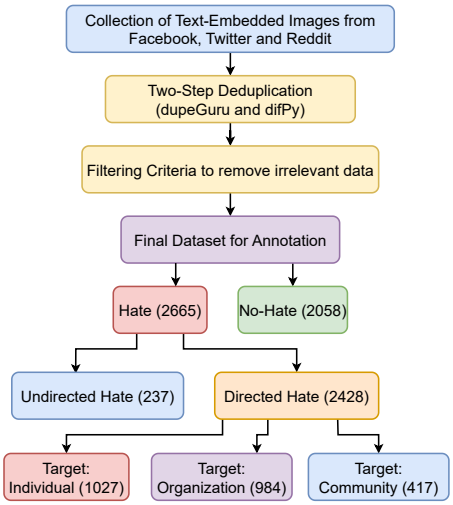


Figure 2. Flow diagram of the data acquisition process

3.1. Data Collection and Deduplication

We collected data from the social media platforms Twitter, Reddit, and Facebook. The Twitter API[1] was used to collect images from Twitter, while manual curation was performed for Reddit and Facebook. We selected several keywords, namely *putin*, *zelensky*, *kremlin*, *ukraine*, *russia*, *kyiv*, *kiev*, *nato*, *russian*, *ukrainian*, *moscow*, *kharkiv*, *donbas*, and *himars* to define our area of interest in data collection. These keywords were chosen to capture relevant text-embedded images of the Russia-Ukraine conflict.

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

For data obtained through Twitter, we used tesseract-OCR[2] to identify text-embedded images which were subjected to manual filtering by using robust filtering criteria mentioned in section 3.2. While collecting images from Facebook and Reddit, the same filtering guideline was used to collect text-embedded images for the curation of the dataset.

Since the data was collected by different individuals from multiple sources, duplicates were encountered. Duplicate data can lead to errors in analysis and negatively impact the quality of results, making it crucial to remove duplicates before analyzing the dataset. We sequentially employed two deduplication tools viz. dupeGuru[3] and Duplicate Image Finder (difPy) python package[4]. Our two-step deduplication helped remove duplicate images by preserving the image with the highest resolution from each batch of duplicates. Google OCR Vision API[5] was used to extract textual content from the images, enabling us to process the data further. By using these tools and techniques, we ensured that the data collection process was clean and consistent.

3.2. Filtering Criteria

Filtering is an essential precursory task to remove data that might potentially skew analysis results. To ensure the relevance and quality of our dataset, we filtered the images based on the following criteria:

- **Non-text or only-text images**: We removed images that did not contain any text or contained only text, such as online articles or images of newspapers.
- **Non-English text**: We manually excluded images that had a considerable amount of non-English words. However, we retained images that had a few commonly used non-English words or phrases such as "Ukraini".
- **Irrelevant Images**: We removed images that were not pertinent to the Russia-Ukraine conflict both visually and textually, such as advertisements, spam, images devoid of context, or images focusing on other unrelated topics.
- **Low-Quality Images**: We eliminated images that had a substantial amount of distortion, blurriness, graininess, or other types of degradation, resulting in incomprehensible text.

Figure 3 shows examples of images that were eliminated during the filtering process. The resulting dataset consists of 4,723 manually annotated images, each with textual and visual content relevant to the Russia-Ukraine conflict.

[2] https://github.com/tesseract-ocr/tesseract
[3] https://github.com/arsenetar/dupeguru
[4] https://github.com/elisemercury/Duplicate-Image-Finder
[5] https://cloud.google.com/vision/docs/ocr

3.3. Annotation Schema

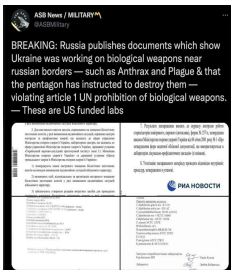
Accurate annotation of data is essential to ensure the consistency, reliability, and validity of the dataset [21]. Furthermore, annotations are also responsible for conveying the underlying patterns within a dataset, which can greatly affect the conclusions drawn from the data. A team of *three annotators* with excellent fluency in the English language annotated the data. The annotators had prior experience annotating and had varying educational qualifications, political beliefs, and backgrounds. Inaccurately or inconsistently labeled data can lead to the distortion of analysis methods and model development. Therefore, we follow a three-phase annotation scheme to ensure that the annotators are well-acquainted with the annotation scheme. To measure the inter-annotator agreement quantitatively, we used Cohen's Kappa (κ) as a measure of the inter-rater agreement. Figure 1 shows examples of annotated images.

3.3.1 3-Phase Annotation

The annotation was initiated with a set of clear and unambiguous instructions. The instructions were further revised in an iterative manner until all annotators were clear about them. As a measure to further eliminate ambiguity and ensure consistency, we followed a three-phase annotation scheme.

- **Pilot Run**: The first phase of annotation involved a pilot run of 50 images to ensure that the annotation instructions were understood by all annotators. Our exhaustively annotated dataset required annotators to have a collective understanding of what constitutes hate speech. During the pilot run, there were slight disagreements among annotators, mainly regarding the targets of hate speech, after which the instructions were revised to address all discrepancies.
- **Revised Instructions**: In the second phase, 200 images were annotated by each annotator to ensure that the revised set of instructions was explicit enough. During this phase, the annotators followed the revised set of instructions to annotate the images. The results of this phase further helped to revise instructions and ensured that annotators were consistently able to identify hate speech.
- **Consolidation Phase**: The third annotation phase involved annotating 50 images in a group meeting by all annotators, during which they discussed discrepancies in their annotation in the second phase and reached a consensus. This phase ensured that all annotations were consistent, helped make instructions more apparent to the annotators, and provided an opportunity to uncover any further ambiguities in the instructions.

(a) Non-text image



(b) Only-text image

(c) Non-English text

(d) Irrelevant Image

(e) Degraded Image

Figure 3. Examples of posts removed during the filtering process. Our robust filtering criteria were applied to remove irrelevant text-embedded images.

3.3.2 Annotation Guidelines

In order to make annotation consistent, we devised detailed annotation guidelines for annotators. The guidelines are mentioned in this section.

Hate Speech: A text-embedded image is classified as hateful if it contains visual or textual hateful content such as threats, personal attacks, slander, or abuse.

- **Targeted language**: Hate speech attributed to the Russia-Ukraine conflict often targets specific groups based on affiliations, political beliefs, or origin. If there is a use of language that degrades, dehumanizes, or demeans a particular organization, a community of people, or an individual, it is labeled as 'Hate Speech'.
- **Hostility and aggression**: If the text-embedded image contains language that promotes hostility, incites aggression, or glorifies violence towards a political organization or individual, it is labeled as 'Hate Speech'.
- **Use of Hateful memes and images**: Hateful memes and images are often used to disseminate harmful language that degrades, dehumanizes, or demeans a particular political organization, a community of people, or an individual.

Hate speech can be masked through sarcasm and satire, making hate more subtle and harder to detect. The annotators were guided to detect sarcasm and satire in images and understand the context in which it is used. They were also trained to differentiate the intent of sarcasm and satire between humor and hate. The annotation guidelines were supplanted with examples of sarcastic and satiric images and how they express hate speech.

No Hate Speech: A text-embedded image is considered non-hateful if it reports events or objectively reports others' opinions in a non-hateful manner. To make guidelines clear, the following points were discussed as significant identifiers of non-hateful speech.

- **Constructive criticism**: Non-hate speech includes constructive criticism of political organizations, policies, parties, and the individuals affiliated with them. This may also include criticism of political events and happenings around the Russia-Ukraine war.
- **Factual and informative**: Non-hate speech includes factual and informative content such as news, reports, updates, and analyses about political proceedings.
- **Respectful and civil**: Non-hate speech during political events remains civil and respectful and does not use hateful symbols or derogatory language.
- **Lack of hostility**: Non-hate speech during the Russia-Ukraine conflict should not express hostility or aggression towards any political group or individual.

Hate speech was further divided into two categories on the basis of direction: 'Directed hate' and 'Non-directed hate'. The annotation guidelines for the direction of hate speech are the following:

- **Directed hate**: Hateful memes and images that are directed towards a particular political organization, a community of people, or an individual.
- **Undirected hate**: Hateful memes and images that do not have a specific target but instead focus on general societal themes or abstract topics, such as war and capitalism, are considered undirected hate. This type of hate speech may also use ambiguous pronouns like "they" or "you" to refer to abstract targets.

Directed hate speech was further divided into three categories based on their intended targets: 'Individual', 'Organization', and 'Community'.

- **Individual**: An individual refers to an autonomous entity involved in politics in any manner. This can include politicians, political candidates, activists, journalists, and other individuals who are involved in political discourse or have a stake in the outcome of the election. Some of the most frequently mentioned individuals in the context of our dataset are "Vladimir Putin", "Volodymyr Zelenskyy", and "Joe Biden".

- **Organization**: An organization refers to a structured group of individuals who come together for a purpose such as a business, non-profit, or government agency. An organization has a clear leadership structure, a specific goal, and a defined membership. Examples of organizations in our dataset include "NATO", "Republican Party", and the "United Nations".
- **Community**: A community refers to a group of people who share a commonality, such as geographical location, culture, or interest. A community does not have a clear leadership structure, and membership is loosely defined. Examples of communities in our dataset include "Russians", "Ukrainians", and "Liberals".

The annotation guidelines were exhaustive, and the annotators regularly communicated the problems in annotations to each other. Some resolutions were made through regular meetings and group annotation sessions. Resolutions were also made through meetings with senior researchers involved in drafting annotation guidelines.

3.3.3 Inter-Annotator Agreement

Table 2. Cohen's Kappa (κ) for annotation during different phases by three annotators

Annotation Phase	Annotators	κ_{taskA}	κ_{taskB}	κ_{taskC}
Pilot Phase	α_1 and α_2	0.64	0.59	0.58
	α_1 and α_3	0.65	0.61	0.60
	α_2 and α_3	0.61	0.54	0.54
Final Phase	α_1 and α_2	0.81	0.77	0.78
	α_1 and α_3	0.82	0.79	0.78
	α_2 and α_3	0.78	0.74	0.72

To assess the consistency of the annotations, Cohen's Kappa (κ) was used as a statistical measure [22]. The inter-annotator agreement was high, with a Cohen's Kappa of 0.78 for Task A, which was a 2-class annotation (κ_{taskA}) of 'Hate' vs 'Non-Hate'. For Task B, a 2-class annotation (κ_{taskB}) of 'Directed' vs 'Undirected', we obtained a Cohen's Kappa of 0.72. Similarly, the Cohen's Kappa for Task C i.e. 3-class annotation (κ_{taskC}) of 'Individual', 'Organization', and 'Community' is 0.71. The annotator agreement for different annotation phases are shown in Table 2.

3.4. Dataset Statistics

Our dataset comprises 4,723 text-embedded images, which have been categorized into two classes: 'Hate' and 'No Hate'. Of these, 2,665 images (56.43%) have been labeled as 'Hate', while 2,058 (43.57%) have been labeled as 'No Hate'. Additionally, the images labeled as hate speech have been further classified into two subcategories, namely, 'Directed hate speech' and 'Undirected Hate Speech'. The 'Directed' category consists of 2,428 (91.11%) images, whereas 'Undirected' consists of 237 (8.89%) images. The directed hate speech was further divided into 'Individual', 'Organization', and 'Community'. The 'Individual' category contains 1,027 images (38.54%), while the 'Organization' category contains 984 images (36.92%). The 'Community' category contains 417 (15.65%). These data figures, along with the average character count and average word count, are presented in Table 3.

Table 3. Statistics for CrisisHateMM. After preprocessing the text embedded in images, the average value of characters per tweet (Avg. Char) and words per tweet (Avg. Words) are determined.

Problem	Labels	Text-embedded images	Avg. Char	Avg. words
Hate Speech	Hate	2,665	200.42 (151.84)	33.34 (28.64)
	Non-Hate	2,058	318.22 (238.70)	51.94 (43.99)
Target	Individual	1,027	194.81 (148.64)	32.39 (28.05)
	Organization	984	201.53 (151.64)	33.45 (28.50)
	Community	417	224.37 (168.22)	37.32 (31.83)
Direction	Directed	2,428	202.67 (153.26)	33.68 (28.89)
	Undirected	237	177.27 (137.32)	29.88 (26.12)

3.5. Exploratory Data Analysis

Table 4 presents the top 10 most frequently occurring words in the hate, direction, and target classes for our dataset. To evaluate the significance of each word in the dataset, we employed the TF-IDF (Term Frequency-Inverse Document Frequency) statistical approach. TF-IDF calculates the weight for each word based on its frequency in a document (TF) and the number of documents in the collection that contain the word (IDF) [23, 24]. The final score of a word is the product of its TF and IDF scores. TF-IDF scores assist in giving useful insights into the patterns and trends that appear in hate speech text. When a word has a high TF-IDF score, it is considered more relevant and meaningful in the context of the document. Table 4 shows that certain words, such as 'Ukraine', 'Russia', and 'Putin', have a high level of significance across most of the subclasses within our dataset. Additionally, Table 4 gives a helpful visual summary of all the words along with TF-IDF scores which may be beneficial in understanding the links between different keywords and the general language used in the dataset.

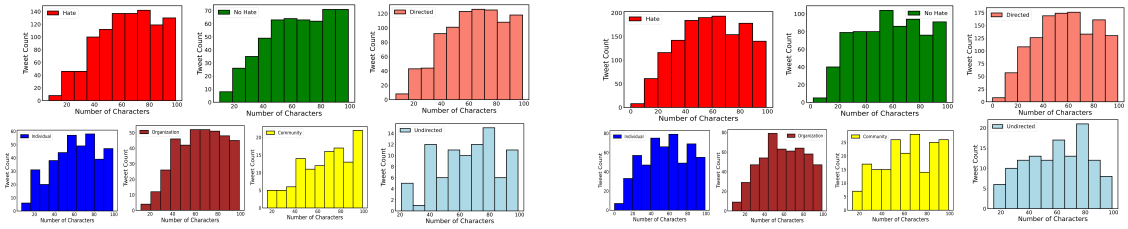
Similarly, the number of characters in each class is shown in a histogram in Figure 4, whereas, Figure 5 shows a histogram of the number of words in each category. Results without text preprocessing as well as the results with text preprocessing are given in the figures. The preprocessing steps are explained in section 4.2.

4. Results and Discussion

We used various techniques to establish the baselines, employing both unimodal and multimodal approaches.
Unimodal methods: In the unimodal process, we utilized various unimodal textual and visual methods:

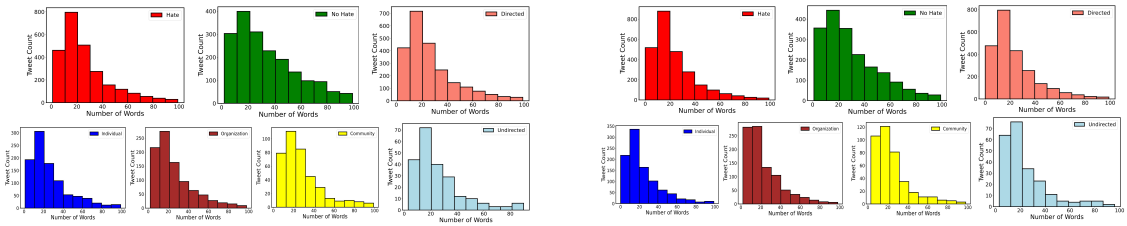
Table 4. Top-10 most frequent words (from text extracted using OCR) in each class. Each word is provided with the TF-IDF scores.

All Posts		Hate Speech Posts		Target: Individual		Target: Organization		Target: Community		Direction: Directed		Direction: Undirected	
Words	**TF-IDF**	**Words**	**TF-IDF**	**Words**	**TF-IDF**	**Words**	**TF-IDF**	**Words**	**TF-IDF**	**Words**	**TF-IDF**	**Words**	**TF-IDF**
ukraine	0.3222	ukraine	0.2863	ukraine	0.2764	ukraine	0.3127	ukraine	0.2508	ukraine	0.2816	ukraine	0.2991
russia	0.2025	russia	0.2139	putin	0.2636	russia	0.3048	russian	0.2165	russia	0.2152	russia	0.1856
russian	0.1448	putin	0.1527	russia	0.1400	russian	0.1471	russia	0.1682	putin	0.1597	war	0.1490
putin	0.1244	russian	0.1377	russian	0.0923	nato	0.0991	ukrainian	0.1344	russian	0.1394	russian	0.1155
war	0.0928	war	0.0899	war	0.0866	war	0.0847	putin	0.0924	war	0.0838	putin	0.0719
ukrainian	0.0915	ukrainian	0.0745	biden	0.0688	ukrainian	0.0699	war	0.0835	ukrainian	0.0768	news	0.0586
news	0.0851	nato	0.0555	president	0.0561	putin	0.0558	people	0.0717	nato	0.0587	ukrainian	0.0545
president	0.0493	world	0.0448	ukrainian	0.0552	world	0.0540	military	0.0394	world	0.0443	world	0.0503
world	0.0475	news	0.0447	vladimir	0.0529	news	0.0419	news	0.0336	biden	0.0436	people	0.0496
nato	0.0472	president	0.0388	news	0.0468	military	0.0396	world	0.0335	news	0.0433	president	0.0292



(a) Without Preprocessing (b) With Preprocessing

Figure 4. Histogram of number of characters per text-embedded image before and after preprocessing



(a) Without Preprocessing (b) With Preprocessing

Figure 5. Histogram of number of words per text-embedded image before and after preprocessing

- **Textual methods**: The textual models used include BERT [25], DistilBERT [26] and DistilRoBERTa [26].
- **Visual methods**: For the visual unimodal baseline methods, we used DenseNet [27], Visformer [28], MViTv2 [29] and VGG19 [30].

Multimodal methods: As we have text-embedded images, we explored using multimodal models to capture the data's visual and textual information. We implemented the state-of-the-art model CLIP (Contrastive Language-Image Pre-training) [31], and GroupViT (Grouping Vision Transformer), a pre-trained vision-language transformer [32].

4.1. Implementation Details

For baselines, we trained the models on a Tesla T4 with 25 GB of dedicated memory and assessed their performance by using accuracy, F1-score (macro), and MMAE (Macro Mean Absolute Error) as performance metrics. We imported the pre-trained transformer models from the hugging face[6] library for the unimodal text and multimodal tasks [33]. Similarly, for the visual models, we employed pre-trained models from the PyTorch Image Models library (timm) [34]. All the tested models, where applicable, used the Adam optimizer [35]. The hyperparameters and models required to replicate the experiments are listed in Table 5.

4.2. Preprocessing

Text preprocessing is an important step in NLP tasks. The text retrieved from OCR was preprocessed along with the image filtering criteria. We removed non-alphanumeric elements, including special characters, hyperlinks, symbols, and non-English characters that may contribute to noise in the data, which could ultimately distort analysis results. Further, non-English words were removed using the English corpus from the NLTK library [36]. Our preprocessing step ensures data quality so that only meaningful text is retained for further analysis.

4.3. Performance Analysis and Insights

Table 6 shows the performance of different algorithms in **Task A**- Hate classification, **Task B**- Direction classification, and **Task C**- Target classification. When using uni-

[6] https://huggingface.co/

Table 5. Implementation Details of the Experiments. Experiments were conducted using a train/test/validation split ratio of 70/15/15.

Modality	Models	Batch Size	Epochs	Learning Rate	Parameters	Image Encoder	Text Encoder
Textual	BERT	8	3	5×10^{-5}	110M	-	bert-base-uncased
	DistilBERT	8	3	5×10^{-5}	67M	-	distilbert-base-uncased
	DistilRoBERTa	8	3	5×10^{-5}	82M	-	distilroberta-base
Visual	DenseNet-161	16	5	10^{-5}	26.5M	densenet161	-
	Visformer_small	16	5	10^{-5}	39.5M	visformer_small	-
	MViTv2_base	16	5	10^{-5}	50.7M	mvitv2_base	-
	VGG19	16	5	10^{-5}	139.6M	vgg19	-
Multimodal	CLIP	4	5	10^{-3}	63M	ViT-Large-Patch14	
	GroupViT	8	5	10^{-3}	-	GroupViT (Hugging Face)	

Table 6. Performance of different unimodal and multimodal algorithms on our dataset

Modality	**Model**	**Hate Classification**			**Direction Classification**			**Target Classification**		
		Accuracy ↑	**F1-score** ↑	**MMAE** ↓	**Accuracy** ↑	**F1-score** ↑	**MMAE** ↓	**Accuracy** ↑	**F1-score** ↑	**MMAE** ↓
Textual	BERT	0.779	0.767	0.240	0.928	0.591	0.427	0.629	0.427	0.998
	DistilBERT	0.754	0.750	0.247	0.925	0.532	0.473	0.637	0.423	1.008
	DistilRoBERTa	0.777	0.769	0.233	0.912	0.578	0.447	0.654	0.440	0.919
Visual	DenseNet-161	0.741	0.739	0.259	0.704	0.487	0.514	0.538	0.425	0.774
	Visformer_small	0.741	0.739	0.257	0.605	0.458	0.461	0.451	0.407	0.772
	MViTv2_base	0.731	0.726	0.276	0.908	0.476	0.500	0.576	0.422	0.657
	VGG19	0.686	0.686	0.305	0.908	0.476	0.500	0.525	0.395	0.785
Multimodal	CLIP	**0.798**	**0.786**	**0.204**	**0.936**	**0.609**	**0.407**	**0.684**	**0.615**	**0.579**
	GroupViT	0.792	0.785	0.214	0.877	0.467	0.500	0.598	0.451	0.763

modal text, DistilRoBERTa performs the best at task A and task C with an F-1 score of 0.769 and 0.440, respectively, despite being a smaller model than BERT. For task B, BERT performs the best with an F-1 score of 0.591. When using unimodal images, DenseNet-161 performs the best with F-1 scores of 0.739, 0.487, and 0.425 for tasks A, B, and C, respectively. For task A, Visformer_small is tied with DenseNet-161 for performance. For unimodal images, the smaller models seemed to perform the best. Among the multimodal models, CLIP outperformed all unimodal and multimodal models, with F-1 scores of 0.786, 0.609, and 0.615 for tasks A, B, and C, respectively. The proportionally higher scores of multimodal models reflect their superiority over unimodal methods, suggesting that multimodal models should be explored more to classify hate speech efficiently. By looking at a few cases of misclassification by models, we can infer that sarcastic images were misclassified. Figure 6 shows an example of such a case.

The misclassification of sarcastic images that pose as harmless memes but carry a more profound meaning through combining visual and textual cues indicates the need to develop multimodal models that better leverage the interplay between visual and textual modalities.

Figure 6. Label: **Hate** Predicted: **No Hate**. Seemingly sarcastic, yet hate speech content is classified as no hate by models.

5. Conclusion

The Russia-Ukraine crisis is a delicate and multifaceted problem that has given rise to many conflicting viewpoints. Identifying hate speech and its targets is essential for protecting prejudiced communities, creating safe spaces, and promoting conflict resolution. In conclusion, this paper presents CrisisHateMM, a novel multimodal dataset of text-embedded images from the Russia-Ukraine conflict annotated for hate and non-hate speech with further sub-classes. Future work could include expanding the CrisisHateMM dataset to include more text-embedded images from other conflicts and social media platforms. Additionally, further research could be conducted to improve the performance of multimodal algorithms for detecting hate speech in text-embedded images.

Acknowledgments

MN acknowledges partial financial support from South Australian DIP – Collaborative Research Funds.

References

[1] Ganesh Chandrasekaran, Tu N Nguyen, and Jude Hemanth D. Multimodal sentimental analysis for social media applications: A comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5):e1415, 2021. 1

[2] DataReportal. *DIGITAL 2023: GLOBAL OVERVIEW REPORT*. Web page: https://datareportal.com/reports/digital-2023-global-overview-report. Accessed March 2023. 1

[3] Anil Singh Parihar, Surendrabikram Thapa, and Sushruti Mishra. Hate speech detection using natural language processing: Applications and challenges. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1302–1308. IEEE, 2021. 1

[4] Sarah T Roberts. *Behind the screen: The hidden digital labor of commercial content moderation*. University of Illinois at Urbana-Champaign, 2014. 1

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

[6] Om Prakash Choudhary, AbdulRahman A Saied, Rezhna Kheder Ali, Sazan Qadir Maulud, et al. Russo-ukrainian war: An unexpected event during the COVID-19 pandemic. *Travel Medicine and Infectious Disease*, 48:102346, 2022. 1

[7] Alexander Shevtsov, Christos Tzagkarakis, Despoina Antonakaki, Polyvios Pratikakis, and Sotiris Ioannidis. Twitter dataset on the Russo-Ukrainian war. *arXiv preprint arXiv:2204.08530*, 2022. 1

[8] Paula Fortuna, Joao Rocha da Silva, Leo Wanner, Sérgio Nunes, et al. A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the third workshop on abusive language online*, pages 94–104, 2019. 2, 3

[9] Juan Carlos Pereira-Kohatsu, Lara Quijano-Sánchez, Federico Liberatore, and Miguel Camacho-Collados. Detecting and monitoring hate speech in Twitter. *Sensors*, 19(21):4654, 2019. 2, 3

[10] Sumit Kumar and Raj Ratn Pranesh. TweetBLM: A hate speech dataset and analysis of black lives matter-related microblogs on Twitter. *arXiv preprint arXiv:2108.12521*, 2021. 2, 3

[11] Lara Grimminger and Roman Klinger. Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. *arXiv preprint arXiv:2103.01664*, 2021. 2, 3

[12] Francesca Gasparini, Giulia Rizzi, Aurora Saibene, and Elisabetta Fersini. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. *Data in brief*, 44:108526, 2022. 2, 3

[13] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33:2611–2624, 2020. 3

[14] Surendrabikram Thapa, Aditya Shah, Farhan Jafri, Usman Naseem, and Imran Razzak. A multi-modal dataset for hate speech detection on social media: Case-study of Russia-Ukraine conflict. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 1–6, 2022. 3

[15] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760, 2017. 2

[16] Fan Yang, Xiaochang Peng, Gargi Ghosh, Reshef Shilon, Hao Ma, Eider Moore, and Goran Predovic. 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, 2019. 2

[17] Chen Liu, Gregor Geigle, Robin Krebs, and Iryna Gurevych. Figmemes: A dataset for figurative language identification in politically-opinionated memes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7069–7086, 2022. 2

[18] Bridget Smart, Joshua Watt, Sara Benedetti, Lewis Mitchell, and Matthew Roughan. #IStandWithPutin versus #IStandWithUkraine: The interaction of bots and humans in discussion of the Russia/Ukraine war. In *Social Informatics: 13th International Conference, SocInfo 2022, Glasgow, UK, October 19–21, 2022, Proceedings*, pages 34–53. Springer, 2022. 2

[19] Mahmud Hasan, Labiba Islam, Ismat Jahan, Sabrina Mannan Meem, and Rashedur M Rahman. Natural Language Processing and sentiment analysis on Bangla social media comments on Russia-Ukraine war using transformers. *Vietnam Journal of Computer Science*, 2023. 2

[20] Cagri Toraman, Oguzhan Ozcelik, Furkan Şahinuç, and Fazli Can. Not good times for lies: Misinformation detection on the Russia-Ukraine war, COVID-19, and refugees, 2022. 3

[21] Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, 2008. 4

[22] Matthijs J Warrens. Five ways to look at cohen's kappa. *Journal of Psychology & Psychotherapy*, 5(4):1, 2015. 6

[23] Sang-Woon Kim and Joon-Min Gil. Research paper classification systems based on TF-IDF and LDA schemes. *Human-centric Computing and Information Sciences*, 9:1–21, 2019. 6

[24] Surabhi Adhikari, Surendrabikram Thapa, Usman Naseem, Priyanka Singh, Huan Huo, Gnana Bharathy, and Mukesh Prasad. 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, 2022. 6

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 7

[26] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 7

[27] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017. 7

[28] Zhengsu Chen, Lingxi Xie, Jianwei Niu, Xuefeng Liu, Longhui Wei, and Qi Tian. Visformer: The vision-friendly transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 589–598, 2021. 7

[29] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. MViTv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022. 7

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 7

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 7

[32] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. GroupViT: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 7

[33] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art Natural Language Processing. *arXiv preprint arXiv:1910.03771*, 2019. 7

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019. 7

[35] Sebastian Bock, Josef Goppold, and Martin Weiß. An improvement of the convergence proof of the adam-optimizer. *arXiv preprint arXiv:1804.10587*, 2018. 7

[36] Edward Loper and Steven Bird. NLTK: The Natural Language Toolkit. *arXiv preprint cs/0205028*, 2002. 7