RUHate-MM: Identification of Hate Speech and Targets using Multimodal Data from Russia-Ukraine Crisis

Surendrabikram Thapa
Virginia Tech
Blacksburg, Virginia, USA
surendrabikram@vt.edu

Farhan Ahmad Jafri
Jamia Millia Islamia
New Delhi, Delhi, India
farhanjafri88888@gmail.com

Kritesh Rauniyar
Delhi Technological University
New Delhi, Delhi, India
rauniyark11@gmail.com

Mehwish Nasim
University of Western Australia
Flinders University
Australia
mehwish.nasim@uwa.edu.au

Usman Naseem
Macquarie University
Sydney, New South Wales, Australia
usman.naseem@mq.edu.au

ABSTRACT

During the conflict between Ukraine and Russia, hate speech targeted toward specific groups was widespread on different social media platforms. With most social platforms allowing multimodal content, the use of multimodal content to express hate speech is widespread on the Internet. Although there has been considerable research in detecting hate speech within unimodal content, the investigation into multimodal content remains insufficient. The limited availability of annotated multimodal datasets further restricts our ability to explore new methods to interpret and identify hate speech and its targets. The availability of annotated datasets for hate speech detection during political events, such as invasions, are even limited. To fill this gap, we introduce a comprehensive multimodal dataset consisting of 20,675 posts related to the Russia-Ukraine crisis, which were manually annotated as either 'Hate Speech' or 'No Hate Speech'. Additionally, we categorize the hate speech data into three targets: 'Individual', 'Organization', and 'Community'. Our benchmarked evaluations show that there is still room for improvement in accurately identifying hate speech and its targets. We hope that the availability of this dataset and the evaluations performed on it will encourage the development of new methods for identifying hate speech and its targets during political events like invasions and wars. The dataset and resources are made available at https://github.com/Farhan-jafri/Russia-Ukraine.

CCS CONCEPTS

• **Information systems** → **World Wide Web**; *Information retrieval*; • **Computing methodologies** → **Natural language processing**.

KEYWORDS

Russia-Ukraine Crisis, Multimodal Data, Content Moderation, Hate Speech

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0172-6/24/05
https://doi.org/10.1145/3589335.3651973

ACM Reference Format:
Surendrabikram Thapa, Farhan Ahmad Jafri, Kritesh Rauniyar, Mehwish Nasim, and Usman Naseem. 2024. RUHate-MM: Identification of Hate Speech and Targets using Multimodal Data from Russia-Ukraine Crisis. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3589335.3651973

1 INTRODUCTION

Social media platforms have revolutionized the way people communicate and express their opinions. This has been a double-edged sword, as while social media platforms have strengthened unique opinions and promoted freedom of speech, they have also served as a fertile ground for cyberbullying, hate speech, and offensive language [22]. The unmatched broadcasting ability of social media platforms has made the dissemination of false information or hate speech a major social concern. As a result, several tools are being developed to help users identify harmful or deceptive content, including everything from sexualization and pornography to hate speech and disinformation [14]. Most of the research is focused on developing sophisticated machine learning algorithms to automatically detect and remove harmful content from their platforms. These algorithms use a variety of features, such as linguistic cues, social context, and user behavior, to identify and flag potentially harmful content [13]. In addition to automated tools, social media platforms are also developing policies and guidelines to govern user behavior and content creation Bhandari et al. [3]. These policies and guidelines are designed to promote responsible and respectful communication on social media platforms, while also protecting users from harmful content. The worldwide prevalence and growing significance of tweets have led to increasing studies on tweets. Several studies have been conducted to investigate the use of hate speech and other forms of offensive language on Twitter, with the aim of identifying strategies to curb their prevalence. Most of the research in tweets today focuses on a few key categories, such as ethnicity, sex, and religion. There are also several types of assaults where the target is mocked and dehumanized [20] that need a good focus of research.

On 24 February 2022, Russia launched a full-scale invasion of Ukraine by land, sea, and air, leading to a polarized response from the international community, with some supporting the invasion and others opposing it [38]. This led to widespread condemnation of

(a) No Hate **(b) Hate** **(c) Tar: Individual** **(d) Tar: Organization** **(e) Tar: Community**

Figure 1: Examples of tweets during Russia-Ukraine Crisis: *Task A*: Hate speech detection - Figure (a - 'No Hate' and b - 'Hate') and *Task B*: Target (Tar) detection - Figure (c - 'Individual', d - 'Organization', and e - 'Community')

the war and sanctions imposed on Russia. As a result, social media became a hub of activity as people expressed their opinions on the humanitarian and economic crisis caused by the invasion. Despite respectful discourse and discussion, there was also a significant amount of hateful content (as shown in Figure 1) targeting various groups of people [5].

Hate speech can have severe consequences for society, and the microblogging and social media platforms put in a significant effort to manage it, mostly through human moderators [21]. However, in political events such as invasions, the volume of hate speech can become overwhelming for human moderators to flag and regulate effectively [3, 31]. Therefore, there is a pressing need for an automated system to identify and manage such content [37]. Thus, to tackle this challenge, we propose a new dataset and benchmark them with multiple state-of-the-art algorithms. Our main contributions are as follows:

- We create and release a new large-scale dataset called **RUHate-MM**, which contains 20,675 manually annotated tweets to identify hate speech and their targets during the Russia and Ukraine crisis.
- We conduct a preliminary analysis of the data. We have set benchmarks with several state-of-the-art textual, visual, and multimodal models.
- Our experiments analysis shows that to address issues of hate speech, an integration of multimodal inputs is important.

The structure of our manuscript is as follows. Section 2 offers an overview of related work. In Section 3, we introduce the dataset, along with annotation guidelines, and present a comprehensive statistical analysis. Section 4 provides detailed information on the experimental methodology. Moving on to Section 5, we present the results and conduct a thorough analysis. Finally, Section 6 serves as the conclusion of the study.

2 RELATED WORKS

In previous studies, different aspects of mischief, such as offensiveness, false news, and hateful tweets, have been investigated separately [22]. Waseem and Hovy [36] annotated 16K tweets-dataset to classify a given tweet as sexist, racist, or neither. The aspect-based annotation is useful for understanding the theme of hate speech. Fortuna et al. [8] introduced a dataset comprising 5,668 tweets that were meticulously annotated across 81 categories of hate speech in the Portuguese language. Concurrently, Pereira-Kohatsu et al. [24] curated a dataset consisting of 6,000 Spanish tweets specifically addressing hate speech. Additionally, they provided an unlabeled corpus containing a substantial 2 million tweets. Recognizing the profound impact of political events on societal perspectives, it becomes imperative to construct datasets that encapsulate a pertinent political context. Kumar and Pranesh [15] addressed this need by presenting TweetBLM, a dataset intricately linked to the Black Lives Matter (BLM) movement. This dataset underwent manual annotation for the identification of hate speech. In a similar vein, Grimminger and Klinger [9] contributed to the landscape by introducing a dataset focused on the 2020 US elections. Their dataset comprised 3,000 tweets, each categorized based on its political stance toward a candidate. Moreover, the tweets were further classified into offensive and non-offensive categories, providing a nuanced understanding of the political discourse surrounding the election.

These studies hold a high importance in studying and understanding hate speech, bullying, and different related tasks. However, these studies focus on unimodal textual data. It is evident that textual data alone cannot provide the information a post wants to convey [2]. Nowadays, most social networks support multimodal content, and therefore it has become extremely important to tackle multimodal hate speech. Kiela et al. [14] proposed multimodal data for detecting hate speech in multimodal memes. Apart from telling if a meme is harmful or not, it is extremely important to identify who is being targeted in the memes. Annotating targets in addition to hate speech can provide valuable insights and help in understanding the nature and scope of hate speech. Sharma et al. [28] developed a meme dataset that contained 3,552 memes about US politics. The annotations of the memes in the dataset include information on whether or not the meme is harmful and the specific group, organization, community, or general public that it targets. The authors also propose DISARM, a multimodal neural architecture for detecting harmful targeting in memes. Similarly, Pramanick et al. [25] presented HarMeme, which is the dataset of 3,544 memes related to the COVID-19 pandemic. The annotation was done for the harm level as well as targets.

Despite the importance of political events, limited work has been done to address hate speech during these events such as the

Russia-Ukraine conflict. Hasan et al. [10] assembled a dataset comprising 10,861 Bengali comments discussing the Russia-Ukraine crisis on YouTube news channels. These comments were systematically categorized into three groups: 'Pro-Ukraine,' 'Pro Russia,' and 'Neutral.' In a parallel effort, Toraman et al. [33] meticulously curated a dataset featuring 5,284 English and 5,064 Turkish tweets. These tweets centered around contemporary issues such as the Russia-Ukraine war, COVID-19, and Refugees, with a specific focus on analyzing the spread of misinformation within the tweeted content. These efforts are important efforts in understanding Russia-Ukraine war sentiments. However, these datasets contain text-only information. The current social media users tend to use image along with text in their tweets. Such posts can propagate hate more easily as they can be more visually appealing than text-only tweets. To address this problem, Thapa et al. [32] presented a data set related to the detection of hate speech during the Russia-Ukraine crisis. They annotated 5,680 text-image pairs related to the crisis which were collected from Twitter. However, the authors only address the problem of hate speech detection but detection of intended targets of hate speech is also an important task which has received limited research interest. Apart from this, authors only annotate around 5,000 tweets which is a small dataset.

The significance of political events, coupled with the harmful consequences of hate speech, underscores the need for the development of hate speech detection tools for such contexts. Hate speech detection and target detection during political events such as the Russia-Ukraine war are necessary to understand the dynamics of hate speech, protect vulnerable groups, develop targeted interventions, and understand and assess the impact of hate speech on society. Furthermore, such analyses hold significant potential for enhancing the understanding of hate speech within the context of wargames, where the dynamics of conflict and the role of information warfare are increasingly complex. Table 1 shows a detailed comparison of previous work in the literature. As seen in the table, our study fills the gap of studying multimodal hate speech and related targets during political events such as the Russia-Ukraine war. To this end, we investigate hateful posts related to the Russia-Ukraine crisis and address two main tasks: (i) Task A: Detecting hateful posts – determining if a post is hateful or not. (ii) Task B: Identification of targets of hateful posts – identifying whether a given hateful post targets an individual, an organization, or a community.

3 DATASET

3.1 Data Collection

The development of the events started when Russia started to attack Ukraine on 22 February 2022. Since we were interested in the tweets around this event, we started to crawl tweets from 22 February 2022 to 28 March 2022 using Twitter API[1]. 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 these keywords mentioned very frequently. The keywords *kharkiv*, and *donbas* were collected from 1 March 2022 as the crisis in those regions started later than 22 February 2023. To filter the data and ensure we collect appropriate data for our study, We included tweets in the dataset that contained media and were written in English. We eliminated tweets that had media in the form of videos or animations. Similarly, If the annotators do not find the tweet to be relevant, it was labeled as 'Non-Informative,' and such tweets were later dropped. The criteria for filtering tweets are given in section 3.2. The resulting dataset comprises 20,675 labeled tweets, each containing an image and text along with annotations. All tweets had unique tweet IDs and hence there are no duplicate data.

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

3.2 Selection of image-text tweets

The tweet (image-text) selection procedure was quite straightforward. The dataset contained a collection of images that included aspects of hate, such as making derogatory comments about specific individuals, targeting a company, and raising issues relevant to the country. The dataset also contained cartoon photos that made fun of politicians, nations, and organizations. Images that were just text-oriented or blurry were not taken into consideration. We were focusing on the English language so the images that majorly included text in other languages such as Russian or Ukrainian were discarded. More concretely, the criteria for dropping data are as follows:

- The tweet contains non-textual elements such as videos, gifs, or audio recordings which do not provide any relevant information about the Russia-Ukraine conflict. We have only considered image as a media.
- The tweet does not have any context. The image and text do not contain any relevant information about the Russia-Ukraine conflict.
- The tweet has a language other than English. Since the tweets revolve around the Russia-Ukraine conflict, there were tweets in Russian and Ukrainian language as well.
- The tweet has a blurry image from which no relevant information can be gained. Such tweets are discarded on a case-by-case basis.
- The tweet contains only a link to an external website or a picture that does not provide any relevant information about the Russia-Ukraine conflict.
- The tweet has more than one image. For the purpose of creating this multimodal dataset, we take tweets with only one image as media to create a text-image pair from a given tweet.

Figure 2 shows some of the images that are filtered. In Figure 2 (a), it can be seen that the text is in the Russian language. Similarly, Figure 2 (b) has a blurred image, and hence no information can be obtained from the tweet. We hence discard such tweets.

3.3 Data Annotation:

Annotation of data was done in two folds, first, as Task A, we annotated the data for binary classes viz. 'Hate' vs 'Non-Hate'. In the second fold, as task B, we annotated the hate speech for target

Table 1: Summary of related datasets in the literature

Work	Year	Data Source	Multimodal	Sub-classes	Size	Context
Waseem and Hovy [36]	2016	Twitter	✗	✗	16,000	✗
Pereira-Kohatsu et al. [24]	2019	Twitter	✗	✗	6,000	✗
Kumar and Pranesh [15]	2021	Twitter	✗	✗	9,165	BLM Movement
Kiela et al. [14]	2020	Self-generated	✓	✗	8,000	✗
Hasan et al. [10]	2023	YouTube	✗	✗	10,861	Russia-Ukraine Crisis
Toraman et al. [33]	2022	Twitter	✗	✗	10,348	Russia-Ukraine Crisis
Grimminger and Klinger [9]	2021	Twitter	✗	✓	3,000	2020 U.S. Election
Thapa et al. [32]	2022	Twitter	✓	✗	5,680	Russia-Ukraine Crisis
Sharma et al. [28]	2022	Online sources	✓	✓	3,552	U.S. Politics
Pramanick et al. [25]	2021	Online sources	✓	✓	3,544	COVID-19
RuHateMM (Ours)	2024	Twitter	✓	✓	20,675	Russia-Ukraine War

(a) Non-English Language (b) Blurred Image (c) Blurred Image (d) No useful Information

Figure 2: Examples of discarded tweet images

labels. For the target labels, we annotated data into three labels viz. 'Community', 'Organization', and 'Individual'. As the name suggests, the class 'Hate' pertains to tweets that are offensive and contains hateful content such as personal attacks, homophobic abuse, racial abuse, or attack on the minority. The 'Non-Hate' class has tweets that report the events objectively and have no offensive or hateful content. Further details on annotation are discussed in Section 3.4. Similarly, for the annotation of targets, we defined targets as the following:

- **Community**: A community is a group of individuals who share common interests, beliefs, or characteristics and interact with one another. It can be defined as a social unit that shares a sense of identity, purpose, and values.
- **Organization**: An organization is a structured group of individuals, created to achieve a specific goal or set of goals. Examples of organizations are 'Republican Party', 'NATO', 'United Nations', etc.
- **Individual**: It refers to a person as an autonomous entity. In the context of our dataset, some of the most frequently noted individuals are 'Putin', 'Trump', 'Biden', etc.

Our team of annotators consisted of four individuals, both male and female, with diverse educational qualifications, including undergraduate, MS, and Ph.D. degrees, as well as researchers with experience in NLP and data collection. All the annotators had at least 10 years of formal English education and were familiar with Russia-Ukraine crisis. This diversity in qualifications helped to ensure clear instructions and maintain a high standard of annotations. The use of diverse annotators is an important aspect of data annotation in order to minimize potential biases [34]. In our study, the diversity in the annotator's background was intended to mitigate any potential bias in the annotation process. The manual annotation process was time-consuming and required significant effort from the annotators, but it helped to ensure a high level of accuracy in the annotations. The use of clear guidelines and regular quality checks were also employed to maintain consistency and reliability in the annotation process.

Three-Phase Annotation: Given the difficulty of labeling tweets that have both text and images, we implemented a 3-phase annotation process as explained in Section 3.4.1.

3.4 Annotation Schema

Clear-cut annotations are crucial in order to ensure that the dataset is labeled consistently and accurately. This is important as the results of any analysis or model development will be based on the labeled data. If the annotations are inconsistent or inaccurate, it can

lead to inaccurate or unreliable results. Clear-cut annotations also ensure that the dataset is representative of the underlying phenomenon being studied and that any conclusions drawn from the data are valid. Thus, we follow a 3-phase annotation. To quantitatively assess the inter-annotator agreement, we have used Cohen's Kappa (κ) (for agreement between two annotators) and Fleiss' Kappa (for agreement between multiple annotators) as our inter-rater agreement measure. We collectively refer to them as Kappa in further sections.

3.4.1 ***3-Phase Annotation***. For annotation of the data, an instruction set was created. The instructions were revised iteratively until all the annotators were clear about the instructions. To ensure that the annotation instructions are clear, we follow a three-phase annotation. The first phase involved a pilot run to make sure everyone understood the annotation. The second phase ensured that the revised instructions after the first phase were clear enough. The third phase eliminated the conflicts in the annotation.

- **Pilot Run:** As the first phase of the annotations, we run a pilot annotation for 50 tweets to ensure that everyone understood the annotation instructions. This is important as the task of labeling tweets can be challenging and it is important that everyone has the same understanding of what constitutes hate speech. There was some confusion among the annotators. Some annotators demanded annotation to be clear and we revised the instructions. The annotation instructions were then revised to address all the confusion.
- **Revised Instructions:** The second phase of annotation of 200 tweets by all four annotators was done to ensure that the instructions revised after the first stage were clear enough. During this phase, the annotators were given the revised instructions and asked to annotate the tweets. This phase was important to confirm that the revised instructions were clear and that the annotators were able to consistently identify hate speech.
- **Conflict Resolution:** The third phase was a group discussion of conflicts in the second phase of annotation, during which the annotators discussed any discrepancies in their annotations and reached a consensus. This phase was important as it helped to resolve any disagreements and ensure that all the tweets were labeled consistently. The group discussion also helped to make the instructions more apparent and provided an opportunity to identify any further ambiguities or inconsistencies in the instructions.

  The annotation guidelines is given in detail in Appendix A.1.

3.5 Inter-Annotator Agreement and Statistics:

To calculate the inter-annotator agreement, we have used Kappa (κ). Kappa is a statistical measure of inter-rater agreement for categorical items. It is a more robust measure of inter-rater agreement than simple percent agreement calculation, as it takes into account the agreement that can be expected by chance [35]. It is used to measure the level of agreement between two or more annotators on a categorical item. The coefficient ranges from -1 to 1, with values close to 1 indicating strong agreement and values close to -1 indicating strong disagreement. A value of 0 represents chance agreement. It's a widely used measure in social science and medical research. For our annotations, the inter-annotator agreement, Fleiss' Kappa for the **Task A** i.e. 2-class annotation of 'Hate' vs 'Non-Hate' ($\kappa_{2-Class}$) is 0.74. Similarly, Fleiss' Kappa for **Task B** i.e. 3-class target annotation ($\kappa_{3-Class}$) is 0.69. The value of Cohen's Kappa among different annotators for different stages of annotation is given in Table 2.

Table 2: Cohen's Kappa (κ) for annotation during different Phases by four annotators

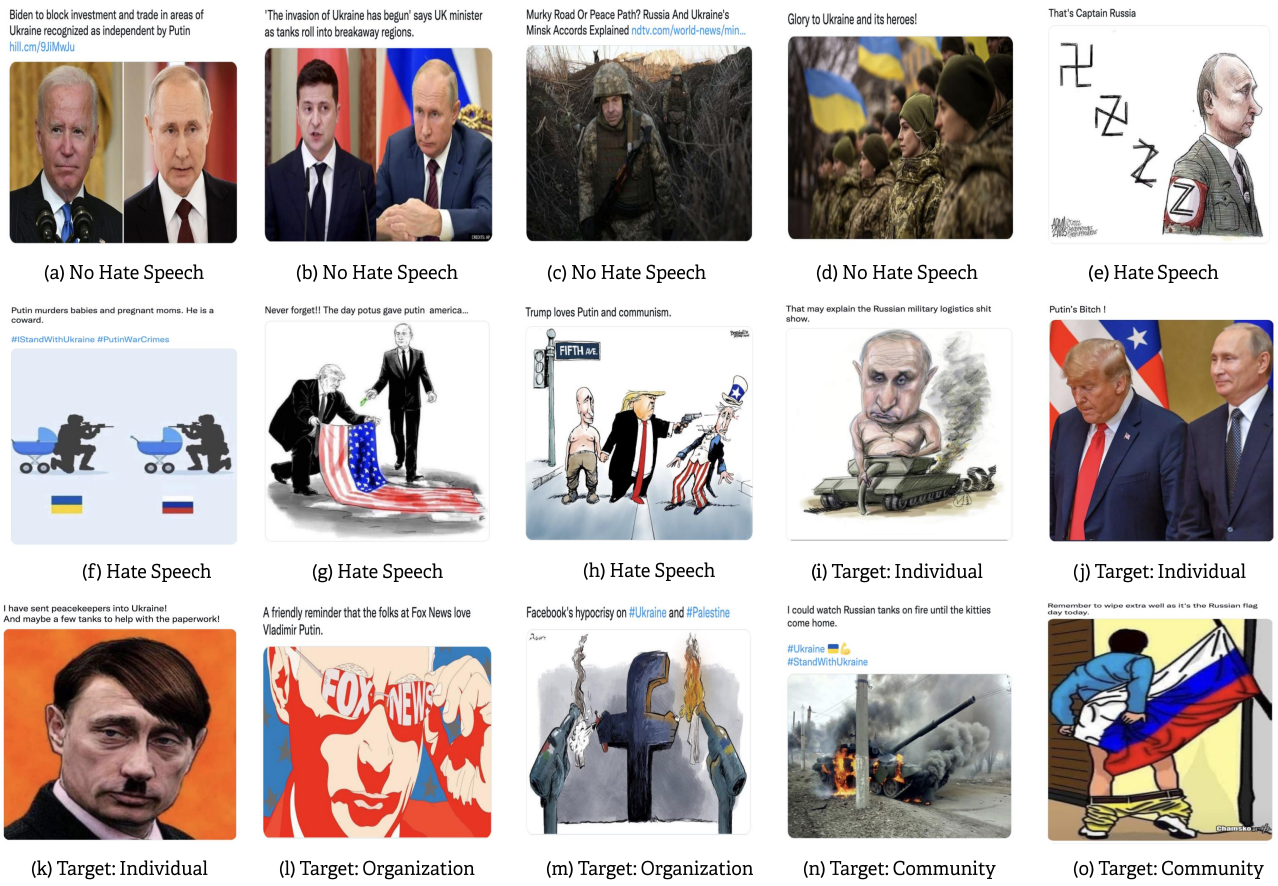
Phase	Annotators	$\kappa_{2-class}$	$\kappa_{3-class}$
Pilot Annotation	α_1 and α_2	0.49	0.53
	α_1 and α_3	0.50	0.51
	α_1 and α_4	0.55	0.56
	α_2 and α_3	0.47	0.52
	α_2 and α_4	0.57	0.60
	α_3 and α_4	0.50	0.49
Final Annotation	α_1 and α_2	0.76	0.74
	α_1 and α_3	0.78	0.76
	α_1 and α_4	0.79	0.76
	α_2 and α_3	0.80	0.73
	α_2 and α_4	0.73	0.71
	α_3 and α_4	0.81	0.73

3.6 Dataset Statistics

Our new multi-modal dataset, RUHate-MM includes 20,675 tweets, with 4,222 (20.33%) tweets being labeled as 'hate speech' label whereas 16,543 (79.67%) tweets are labeled as 'no hate' label (Table 3). The dataset statistics represent a true distribution in a real-world scenario where many posts are neutral, and only some are related to hate speech. The hate speech tweets were further annotated into three broad categories of targets. The targets are individuals, organizations, and communities. Among 4,222 hate speech tweets, 2,402 targeted individuals, 918 of them targeted organizations whereas the remaining 902 posts targeted specific communities. The statistics of data are shown in table 3. It also provides the average number of characters and words per tweet for each of the categories. These statistics can be used to gain an understanding of the overall composition and size of the dataset. The values in parentheses are calculated after preprocessing of text, which includes cleaning and normalizing the text data. Further details of text preprocessing can be found in Section 4.3. It can further be observed that the average character counts for hate and non-hate speech is somewhat distinctive. Figure 1 shows examples of annotated tweets whereas Figure 3 shows some more examples of targeted hate speech.

3.7 Exploratory Data Analysis

Table 4 presents the top 10 most frequent words in each class of the dataset along with their TF-IDF scores. TF-IDF (term frequency-inverse document frequency) is a statistical measure that is used to



(a) No Hate Speech (b) No Hate Speech (c) No Hate Speech (d) No Hate Speech (e) Hate Speech

(f) Hate Speech (g) Hate Speech (h) Hate Speech (i) Target: Individual (j) Target: Individual

(k) Target: Individual (l) Target: Organization (m) Target: Organization (n) Target: Community (o) Target: Community

Figure 3: A few examples from our dataset. The tweets are annotated as hate and non-hate for *Task A*. Hate tweets are further annotated into targets (Tar) as an individual, organization, and community for *Task B*

Table 3: Dataset Statistics for RUHate-MM. The values in parentheses in average characters per tweet (Avg. Char) and average words per tweet (Avg. words) are calculated after preprocessing of text.

Problem	Labels	Tweets	Avg. Char	Avg. words
Hate Speech	Hate	4,222	54.56 (46.46)	9.94 (9.63)
	Non-Hate	16,543	56.14 (51.63)	10.22 (10.12)
Targets	Individual	2,402	53.80 (45.62)	9.87 (9.57)
	Organization	918	56.85 (48.31)	10.14 (9.84)
	Community	902	54.23 (46.79)	9.92 (9.58)

evaluate the importance of a word in a document or a collection of documents. The fundamental principle behind TF-IDF involves assigning weights to words based on their frequency within a document and their rarity across the entire corpus. The TF-IDF score for a word in a document is calculated as the product of its term frequency (TF) and inverse document frequency (IDF) scores. Words with higher TF-IDF scores are deemed more significant and relevant to the document in which they appear. This significance arises from their frequency within the document and their scarcity across the broader collection of documents.

Analyzing the presented table 4, it becomes evident that certain prominent personalities, such as Putin, Trump, Biden, etc., are consistently targeted more frequently in individual categories. TF-IDF aids in highlighting these key terms by emphasizing their importance within specific documents while considering their rarity across the entire dataset. This approach enables a more nuanced understanding of the distinctive features and themes associated with each class, shedding light on the salient words that contribute significantly to the characterization of each category.

4 EXPERIMENTAL SETUP

In this section, we describe our experimental setup and baselines.

4.1 Experimental Settings

We used the pre-trained models for each baseline and used the F1-score as the evaluation metric. We train all the models using Tensorflow and Pytorch on a Tesla T4 or Tesla V100 GPU, with 32

Table 4: Top-10 most frequent words in each class. The TF-IDF scores are given for each words.

All Posts		Hate Speech Posts		HS: Target Individual		HS: Target Organization		HS: Target Community	
Words	TF-IDF	Words	TF-IDF	Words	TF-IDF	Words	TF-IDF	Words	TF-IDF
ukraine	0.3463	putin	0.3274	putin	0.4524	ukraine	0.2284	ukraine	0.2997
russia	0.1659	ukraine	0.2025	ukraine	0.1449	putin	0.1684	russia	0.2161
putin	0.1629	russia	0.1458	russia	0.1140	russia	0.1531	russian	0.1883
russian	0.1211	russian	0.1320	russian	0.0931	russian	0.1529	putin	0.1030
ukrainian	0.0812	trump	0.0488	trump	0.0746	nato	0.1027	ukrainian	0.0694
war	0.0505	war	0.0459	biden	0.0433	ukrainian	0.0572	war	0.0587
kyiv	0.0333	ukrainian	0.0429	war	0.0406	nazis	0.0554	fuck	0.0333
people	0.0245	nato	0.0305	like	0.0259	war	0.0524	people	0.0323
says	0.0241	biden	0.0287	ukrainian	0.0242	nazi	0.0438	just	0.0238
biden	0.0216	just	0.0241	fuck	0.0234	nestle	0.0191	embassy	0.0186

GB dedicated memory. For the unimodal text models, we import all the pre-trained models of transformers from the hugging-face library. Similarly, for the visual models, we import all the pre-trained models from the PyTorch Image Models library (timm) [23]. All the models we experimented with used Adam optimizer [4]. The hyper-parameters that can be supplemental to reproduce the experiments are given in Table 5.

4.2 Baselines

We established baselines using various techniques, including both unimodal and multimodal methods.

Unimodal Models: We used the following unimodal methods:

- **Textual Unimodal:** For textual models, we used Bidirectional Encoder Representations from Transformers (BERT) [7], DistilBERT (a distilled version of BERT) [27], optimized variant of BERT, i.e., RoBERTa [18] and Albert (A Lite BERT for Self-supervised Learning of Language Representations) [16].
- **Visual Unimodal:** For the image-based unimodal baseline methods, we used 4 pretrained methods viz. DenseNet [12], Visformer [6], Improved Multiscale Vision Transformers for Classification and Detection (MVITV2) [17] and VGG19 [29].

Multimodal Models: We employed 3 multimodal models that have been widely used in earlier research involving hate speech classification. The first is a combination of ResNet [11] and BERT, where we first trained ResNet and BERT on text and image data respectively, and then merged their representations using a linear layer. We also implemented the state-of-the-art model CLIP (Contrastive Language-Image Pre-training) in our study, which has shown remarkable performance in a wide range of vision-and-language tasks [26]. We also utilized ViLBERT (Visual-Linguistic BERT) in our study, which is a pre-trained transformer-based model that is specifically designed to handle vision-and-language tasks [19].

4.3 Text Preprocessing

Text preprocessing is important in any Natural Language Processing (NLP) task [1, 30]. In order to prepare the tweet text for further analysis, a preprocessing step was conducted to remove various non-alphanumeric elements, including but not limited to special characters, hyperlinks, mentions, and emojis. Special characters, such as punctuation marks and other symbols, were removed as they can add noise to the data and potentially impact the accuracy of subsequent analysis. Hyperlinks and mentions, which reference external web pages or other users respectively, were removed as they are not relevant to the content of the tweet itself and can also introduce noise into the data. Emojis, while commonly used in social media, were also removed as they are not part of the standard character set used in most natural language processing techniques, and thus could cause errors or inaccuracies in downstream analysis. Overall, the preprocessing step serves to clean and standardize the text data for further analysis, while ensuring that only relevant and meaningful content is retained.

5 RESULTS AND ANALYSIS

In this section, we report the performance of various models on both tasks viz. hate speech classification and target classification.

5.1 Unimodal Baseline Results

Table 6 summarizes the performance of different unimodal algorithms on our dataset in terms of $F1_{HateSpeech}$ (f1-score in hate speech detection) and $F1_{Target}$ (f1-score in target identification). The results for the unimodal textual models (BERT, DistilBERT, ROBERTa, and Albert) demonstrate competitive performance in hate speech detection, with F1 scores ranging from 0.749 to 0.798. Among these, BERT exhibits the highest F1 score, indicating its effectiveness in identifying hate speech instances.

In target identification, the unimodal textual models also showcase notable performance, with $F1_{Target}$ scores ranging from 0.644 to 0.679. BERT again emerges as the leading model in target identification.

For unimodal visual models (DenseNet-161, Visformer_small, MVITV2_base, and VGG19), the results in hate speech detection (ranging from 0.758 to 0.774) and target identification (ranging from 0.598 to 0.628) demonstrate a competitive performance across the board. DenseNet-161 achieves the highest F1-score in both hate speech detection and target identification among the visual models.

Table 5: Implementation Details of the Experiments

Modality	Models	Batch Size	Epochs	Learning Rate	Parameters	Image Encoder	Text Encoder
Textual	BERT	16	3	5×10^{-5}	110M	-	bert-base-uncased
	DistilBERT	16	3	5×10^{-5}	67M	-	distilbert-base-uncased
	RoBERTa	16	3	5×10^{-5}	125M	-	roberta-base
	ALBERT	16	3	5×10^{-5}	12M	-	albert-base-v2
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	ResNet + BERT	16	5	10^{-3}	172M	ResNet-152	Bert-base-uncased
	CLIP	4	6	10^{-3}	63M	ViT-Large-Patch14	
	ViLBERT-CC	16	5	10^{-3}	112M	FasterRCNN	Bert-base-uncased

5.2 Multimodal Baseline Results

The performance of multimodal algorithms on our dataset is presented in Table 6, focusing on $F1_{HateSpeech}$ and $F1_{Target}$. The multimodal models, combining ResNet with BERT, CLIP, and ViLBERT-CC, exhibit enhanced performance compared to unimodal models.

In hate speech detection, the multimodal models achieve higher $F1_{HateSpeech}$ scores compared to their unimodal counterparts. Particularly, ViLBERT-CC stands out with the highest $F1_{HateSpeech}$ score of 0.848, indicating its superior ability in identifying hate speech instances when leveraging both textual and visual modalities.

Similarly, in target identification, multimodal models outperform unimodal models, emphasizing the effectiveness of leveraging both textual and visual information. ViLBERT-CC leads in $F1_{Target}$ with a score of 0.741, demonstrating its proficiency in identifying diverse targets within hate speech instances.

These results underscore the advantages of multimodal approaches, showcasing improved performance in hate speech detection and target identification compared to unimodal models on our dataset. The combination of textual and visual information proves to be a promising avenue for advancing the field of hate speech detection during politically charged events.

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

Modality	Model	$F1_{HateSpeech}$	$F1_{Target}$
Unimodal Textual	BERT	0.798	0.679
	DistilBert	0.787	0.668
	ROBERTa	0.780	0.668
	Albert	0.749	0.644
Unimodal Visual	DenseNet-161	0.768	0.628
	Visformer_small	0.762	0.611
	MVITV2_base	0.774	0.625
	VGG19	0.758	0.598
Multimodal	ResNet + BERT	0.806	0.700
	CLIP	0.826	0.719
	ViLBERT-CC	0.848	0.741

5.3 Analysis

Multimodal models, leveraging both textual and visual information, consistently outperform unimodal models in hate speech detection and target identification. ViLBERT-CC particularly shines, showcasing superior performance and emphasizing the synergistic benefits of combining both modalities. This suggests a promising avenue for future research, highlighting the potential of multimodal approaches in enhancing the discernment of hate speech instances and the identification of diverse targets during complex socio-political events.

The benchmarked evaluations reveal areas for improvement in accurately identifying hate speech and its targets, indicating the intricate nature of the dataset. This opens avenues for refining existing models and developing novel approaches to address the unique challenges presented by hate speech during geopolitical crises. The release of our multimodal dataset contributes a valuable resource to the research community, fostering further exploration and innovation in hate speech detection. The dataset's annotations, particularly the categorization of hate speech targets, offer a nuanced understanding of the manifestations of hate speech during the Russia-Ukraine crisis.

6 CONCLUSION

This paper introduces a new multi-modal dataset for identifying hateful content on social media, consisting of 20,675 text-image pairs collected from Twitter, labeled for hate speech and their targets. The experimental analysis of the presented dataset has shown that understanding both text and image modalities is crucial for detecting hateful content. In future work, we plan to develop new multi-modal models specifically for hate-speech detection, with the goal of gaining a deeper understanding of the relationship between text and images. Another potential area of research could be expanding the dataset to include more languages and different types of social media platforms. Additionally, it would be interesting to explore the hate speech detection possibilities for more specific or nuanced targets.

REFERENCES

[1] 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 (2022), 102761.

[2] Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022. A Survey on Multimodal Disinformation Detection. In *Proceedings of the 29th International Conference on Computational Linguistics*. 6625–6643.

[3] 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*. 1993–2002.

[4] Sebastian Bock, Josef Goppold, and Martin Weiß. 2018. An improvement of the convergence proof of the ADAM-Optimizer. *arXiv preprint arXiv:1804.10587* (2018).

[5] Jean-Christophe Boucher. 2022. Disinformation and Russia-Ukrainian War on Canadian Social Media. *The School of Public Policy Publications* 15, 1 (2022).

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

[7] 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* (2018).

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

[9] Lara Grimminger and Roman Klinger. 2021. Hate Towards the Political Opponent: A Twitter Corpus Study of the 2020 US Elections on the Basis of Offensive Speech and Stance Detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. 171–180.

[10] Mahmud Hasan, Labiba Islam, Ismat Jahan, Sabrina Mannan Meem, and Rashedur M Rahman. 2023. Natural Language Processing and Sentiment Analysis on Bangla Social Media Comments on Russia–Ukraine War Using Transformers. *Vietnam Journal of Computer Science* (2023), 1–28.

[11] 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*. 770–778.

[12] 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*. 4700–4708.

[13] Farhan Ahmad Jafri, Mohammad Aman Siddiqui, Surendrabikram Thapa, Kritesh Rauniyar, Usman Naseem, and Imran Razzak. 2023. Uncovering Political Hate Speech During Indian Election Campaign: A New Low-Resource Dataset and Baselines. (2023).

[14] 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 (2020), 2611–2624.

[15] Sumit Kumar and Raj Ratn Pranesh. 2021. Tweetblm: A hate speech dataset and analysis of black lives matter-related microblogs on twitter. *arXiv preprint arXiv:2108.12521* (2021).

[16] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).

[17] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. 2022. MViTv2: Improved Multiscale Vision Transformers for Classification and Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4804–4814.

[18] 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* (2019).

[19] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* 32 (2019).

[20] Lambert Mathias, Shaoliang Nie, Aida Mostafazadeh Davani, Douwe Kiela, Vinodkumar Prabhakaran, Bertie Vidgen, and Zeerak Waseem. 2021. Findings of the WOAH 5 shared task on fine grained hateful memes detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*. 201–206.

[21] Sünje Paasch-Colberg and Christian Strippel. 2022. "The Boundaries are Blurry…": How Comment Moderators in Germany See and Respond to Hate Comments. *Journalism Studies* 23, 2 (2022), 224–244.

[22] 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)*. IEEE, 1302–1308.

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

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

[25] Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md Akhtar, Preslav Nakov, Tanmoy Chakraborty, et al. 2021. Detecting harmful memes and their targets. *arXiv preprint arXiv:2110.00413* (2021).

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

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

[28] Shivam Sharma, Md Akhtar, Preslav Nakov, Tanmoy Chakraborty, et al. 2022. DISARM: Detecting the victims targeted by harmful memes. *arXiv preprint arXiv:2205.05738* (2022).

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

[30] Surendrabikram Thapa, Surabhi Adhikari, Usman Naseem, Priyanka Singh, Gnana Bharathy, and Mukesh Prasad. 2020. Detecting Alzheimer's disease by exploiting linguistic information from Nepali transcript. In *Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 18–22, 2020, Proceedings, Part IV 27*. Springer, 176–184.

[31] Surendrabikram Thapa, Farhan Jafri, Ali Hürriyetoğlu, Francielle Vargas, Roy Ka-Wei Lee, and Usman Naseem. 2023. Multimodal hate speech event detection-shared task 4, case 2023. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text*. 151–159.

[32] 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.

[33] Cagri Toraman, Oguzhan Ozcelik, Furkan Şahinuç, and Fazli Can. 2022. Not Good Times for Lies: Misinformation Detection on the Russia-Ukraine War, COVID-19, and Refugees. *arXiv preprint arXiv:2210.05401* (2022).

[34] 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*. 7174–7183.

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

[36] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*. 88–93.

[37] P William, Ritik Gade, Rup esh Chaudhari, AB Pawar, and MA Jawale. 2022. Machine Learning based Automatic Hate Speech Recognition System. In *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*. IEEE, 315–318.

[38] Xi-Yin Zhou, Gang Lu, Zhicheng Xu, Xiaoqing Yan, Soon-Thiam Khu, Junfeng Yang, and Jian Zhao. 2023. Influence of Russia-Ukraine War on the Global Energy and Food Security. *Resources, Conservation and Recycling* 188 (2023), 106657.

A APPENDIX

A.1 Annotation Guidelines

A.1.1 Hate Speech. A post (text or image or both) that contains hateful content such as a personal attack, homophobic abuse, racial abuse, or attack on minorities.

- **Targeted language:** Hate speech during political events often targets specific groups based on their political beliefs or affiliations. This can include language that demeans, degrades, or dehumanizes a particular political group or individual.

- **Hostility and aggression:** Hate speech during political events often expresses hostility or aggression towards a particular political group or individual. This can include language that promotes or glorifies violence or hatred against a particular political group or individual.
- **Use of memes and images:** Hate speech during political events often uses memes and images to disseminate harmful messages that are intended to demean, degrade or dehumanize a particular political group or individual.

Further, it is important to note that sarcasm and political satire can be used to express hate speech, and they can be difficult to identify. Sarcasm and satire can be used to mask hate speech, making it more subtle and harder to detect. Sarcasm can be used to express hate speech in a way that is less obvious and less likely to be flagged as hate speech. Satire can also be used to express hate speech in a way that is intended to be humorous or satirical but can still be harmful. Annotators were trained to identify sarcasm and satire in tweets and understand the context in which it is used. They were also trained to differentiate between sarcasm and satire which is intended to be humorous or satirical and sarcasm and satire which is used to express hate speech. Annotation guidelines included clear examples of sarcasm and satire and how they can be used to express hate speech.

A.1.2 No Hate Speech. A post (text or image or both) reports the events or others' opinions objectively and contains no offensive or hateful content. To make guidelines clear, the following were discussed as the significant characteristics of non-hate speech.

- **Constructive criticism:** Non-hate speech during political events often includes constructive criticism of political figures, policies or parties. It can also include criticism of political events and happenings.
- **Factual and informative:** Non-hate speech on Twitter during political events often includes factual and informative content, it can be news, updates, and analysis of the political events.
- **Respectful and civil:** Non-hate speech during political events is respectful and civil in nature, it doesn't use derogatory language or hate speech symbols.
- **Lack of hostility:** Non-hate speech during political events does not express hostility or aggression towards a particular political group or individual.
- **Lack of misinformation or fake news:** Non-hate speech during political events does not spread misinformation or fake news, it is based on facts and credible sources.
- **Lack of targeting specific group:** Non-hate speech during political events does not target specific groups of people based on their political beliefs or affiliations.
- **Lack of sarcasm and/or political satire intended for hate:** Non-hate speech during political events does not use sarcasm or political satire to express hate speech.

A.1.3 Conflict Resolution Between Organization and Community Targets. Organization refers to a group of people who come together for a specific purpose, such as a business, non-profit, or government agency. An organization can have a clear leadership structure, a specific goal or mission, and a defined membership. Community refers to a group of people who share a common bond, such as geographic location, culture, or interest. A community can be more loosely defined than an organization, and it may not have a clear leadership structure or defined membership.

The guidelines for delineating organization and community are given below:

Community Targets:

- The tweet references a specific group of people based on their shared characteristics, such as race, ethnicity, national origin, religion, sexual orientation, gender identity, or disability with an intention of demeaning.
- The tweet references a specific geographic location or culture or specific interest or shared activity as the basis for the targeted group with an intention of demeaning it.

Organization Targets:

- The tweet specifically mentions the name of an organization or a business.
- The tweet references a specific goal or mission of the organization. The tweet references a specific set of actions or decisions made by the organization.
- The tweet references a specific leadership structure or hierarchy within the organization.

The annotation guidelines were exhaustive and the annotators regularly communicated the problems in annotations to each other. Some resolutions in annotation were made through meetings and annotation sessions. Annotators were able to distinguish between tweets that target an organization and tweets that target a community by analyzing the language used in the tweets and understanding the context in which the tweets were written.

A.2 Limitations

In this paper, we present a large-scale multimodal dataset for hate speech detection and target identification. We also present baselines for detecting hate speech and identifying targets using this dataset. However, there are several limitations to our work that should be acknowledged. First, our dataset is limited to tweets from a specific time period surrounding a political event, and may not be representative of hate speech in other contexts. Additionally, our dataset is based on tweets from a single microblogging platform, and it is not clear how well our approach would generalize to other platforms or modalities. Second, our annotation scheme for targets is based on broad categories (Individuals, Organizations, and Communities), and may not capture more specific or nuanced targets. Furthermore, the annotation process is subjective, and different annotators may have different opinions on whether certain tweets should be considered hate speech or not. Third, the baselines we provide are based on a limited set of features, and it is possible that other features or architectures could lead to improved performance. Finally, it's important to note that hate speech detection and target identification technologies can raise ethical concerns, such as the potential for misuse, bias, and invasion of privacy. These ethical concerns should be considered and addressed in the development and deployment of such technology.