"Stronger than Hate": On the Dissemination of Hate Speech during the 2020 Vienna Terrorist Attack

Michaela Lindenmayr*, Ema Kušen*, Mark Strembeck*†‡
*Vienna University of Economics and Business, Austria
†Complexity Science Hub, Vienna, Austria
‡Secure Business Austria (SBA), Vienna, Austria
E-mail: {firstname.lastname}@wu.ac.at

Abstract—**In this paper, we present an analysis of 36,685 tweets related to the 2020 Vienna terror attack. We used a Convolutional Neural Network (CNN) approach to identify hateful and non-hateful tweets. Our findings indicate that users who post hateful content are predominantly anonymous. Moreover, we found that hateful messages can spread widely across the network and that hateful communication forms characteristic structural patterns.**

Index Terms—**Hate Speech, Machine Learning, Network Motifs, Neural Network, Twitter, Terrorist Attack**

I. Introduction

On November 2^{nd} 2020 at 8:00pm, the Vienna police received an emergency call about a man firing shots and attacking by-passers in a restaurant district close to the inner city. The police soon arrived on the scene and the attacker was shot dead at 8:09pm while the police continued searching for other perpetrators. It was later confirmed that a single attacker, a 20-year old ISIS-supporter who was born and raised in Austria, was responsible for the attack. Before being stopped by the police, the attacker killed four people and injured 23 more. In the aftermath, many others have been struggling with psychological trauma. In the following days, the police extended their search for the perpetrator's network and arrested suspects in Austria and Germany.

In this paper, we analyze a data-set consisting of 36,685 tweets related to the 2020 Vienna terror attack. We investigate the dissemination of hateful messages on Twitter during the terror attack and in its immediate aftermath. In particular, we analyze the communication structures (network motifs) that arise from direct messaging on Twitter and their temporal characteristics.

While there is the right for freedom of speech (even if the content of a message is hateful), hate speech is considered dangerous [20] because of its frequent occurrence [24] and because of the significant impact such messages have on those who are targeted by it [8].

For the purposes of our analysis, we follow a definition which describes *hate speech* as content that aims to seriously disparage or attack individuals or groups based on some protected characteristics which include race, color, ethnicity, gender, sexual orientation, nationality, and religion (see [13], [17], [22], [28], [30], [35], [39]).

The remainder of this paper is organized as follows. In Section II, we give an overview of related work. Section III outlines the research procedure before our findings are presented in Section IV and discussed in Section V. Section VI concludes the paper.

II. Related Work

A. Hate speech detection approaches

Profane words are one of the main indicators of hate speech, while so called cyber-bullying words might further improve the detection of hate speech [10].

One approach for hate speech detection relies on the Bag-Of-Words (BOW) technique that uses a pre-defined word corpus [5], [7]. However, [7] found that this approach leads to a low accuracy and a high false positive rate. To address this limitation, [27], [7], and [31] suggested the use of n-grams. In [18], the most accurate classification was reached by using a combination of word unigrams, Part-Of-Speech (POS) tagging, and emoji features.

When it comes to machine learning models that predict whether a message is hateful or not, [31] evaluated logistic regression, Support Vector Machines (SVM), random forests, and gradient boosting. They found that the SVM performed best with a true positive rate of 89.4%. Similar accuracy was also reported in [2] where SVMs in combination with term frequency–inverse document frequency (TF-IDF), char quad-grams, word unigrams, number of positive and offensive words, and the proportion of positive and offensive words reached an F1-measure of 85% . Moreover, [19] found that SVMs are able to outperform the logistic regression approach. A study performed by Ibrohim et al. [18] compared logistic regression, random forest decision trees, and SVMs. In their study, the logistic regression was found to outperform the other approaches.

Besides the approaches mentioned above, deep learning models have also been used to identify hate speech. For example, [19] uses word embeddings for pre-training models before testing the performance of traditional machine learning models, i.e., SVM and logistic regression, against deep neural network models (a Convolutional Neural Network (CNN) and a Gated Recurrent Unit Neural Network (GRU)). While the SVM outperformed the logistic regression and reached an accuracy of 65.10%, the CNN reached an accuracy of 83.04%. Badjatiya et al. [3] found that neural network approaches perform better than SVMs, logistic regression, and Gradient

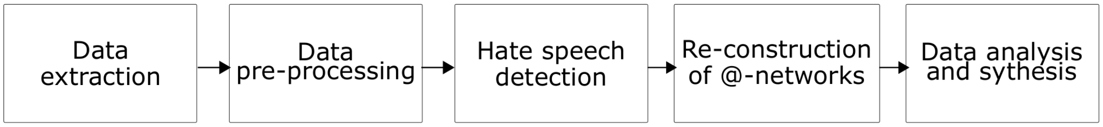

Fig. 1. Research procedure.

Boosting Decision Trees, while they consider CNNs to be the best approach in their study.

B. Characteristics of hate speech and its spreaders

Dadvar et al. [10] found that users who frequently published hateful content in the past are also more likely to publish hateful content in the future when compared to other users. In [1], Arango et al. found that the followers of a user, his/her subscriptions, as well as likes serve as good indicators of the likelihood of publishing hateful content. Another study [9] reported that juveniles between 13 and 16 are most likely to spread hateful content on YouTube, and that male users are more likely to post hateful or offensive content than female users. Such users also mostly try to hide their real identity [11].

Chakraborty et al. [6] found that the tweets authored by abusive users are re-posted with a lower likelihood and that the abusive users tag their friends less frequently in their messages. In addition, abusive messages include more URLs and hashtags than non-abusive tweets, and tweets of the same (abusive) user have a similar content [6]. Chetty and Alathur [8] report that the proportion of people responsible for hate speech on platforms is low, while the impact of those few people can be considerable.

Intentional misspelling, the use of informal abbreviations, or short forms of a word are frequently used to avoid the detection by standard filters against offensive language [11], [32], [37]. Moreover, users who spread hateful content are more likely to post shorter comments compared to other users [6].

III. Research procedure

For our analysis, we defined the following research questions (RQ):

- **RQ1: What are the characteristics of hate speech on Twitter during the 2020 Vienna terror attack?** This RQ examines the temporal flow of hateful messages, their content, as well as user reactions to hate speech.
- **RQ2: What are the characteristics of the main spreaders of hate?** In this RQ, we focus on the anonymity of the message spreaders and their content creation (tweeting) rate.
- **RQ3: How does hate disseminate over Twitter?** This RQ examines the retweeting of hateful messages, as well as the properties of the resulting hate-exchange network and the characteristic patterns (network motifs) that emerge in the respective communication networks.

Our research procedure includes five steps (see Figure 1).

Data extraction. We used Twitter's Search API to extract tweets written in English language using a set of key-words and hashtags related to the event. The key-words and hashtags we used are *"#viennaterror", "viennaterror", "#viennaterrorattack", "viennaterrorattack", "#viennaterrorist", "viennaterrorist", "#terroristattackvienna", "terroristattackvienna", "#viennaattack", "viennaattack"*. The data extraction started on the November 2 and stopped on November 11.

Data pre-processing. After collecting the data, we removed tweets that only contain URLs, resulting in a data-set including 36,685 unique tweets. We also cleaned up redundant blank spaces, carriage returns, line feeds, and stop words. In addition, we stemmed the words.

Hate speech detection. We used three techniques for hate speech detection: 1) a lexicon approach, 2) a support vector machine (SVM) classifier, and 3) a deep learning approach.

1) The lexicon approach uses a compilation of existing dictionaries, including a list of words banned by Google [14], a list of swear words published by NoSwearing [29], a list of offensive words published by the Carnegie Mellon University [36], a data-set from Davidson et al. adapted from Hatebase [12], a data-set on hate speech from Mathew et al. [21], and the Hatebase list of words [16]. Moreover, we adapted the dictionaries by removing the terms "attack", "terror", "terrorist", "dead", "kill", and "killed" because in our case these terms are regarded as general terms and might therefore lead to an incorrect classification of hate speech. Prior to applying a lexicon approach, we stemmed the lexicon entries. The final list of entries counts 2,483 stemmed terms.
2) For the SVM approach, we used Python scikit-learn for deriving the SVM labels. We considered the following features to train our classifier – BOW, TF-IDF approach for unigrams, bigrams, trigrams, and POS tags. We trained the classifier with a linear and sigmoid kernel but only report on the accuracy scores for the sigmoid kernel since it achieved a higher accuracy.
3) For the Convolutional Neural Network (CNN), we used the Tensorflow package for R [33], with random embeddings as well as GloVe embeddings as input features. One convolutional layer of 512 filters with kernel size 3 was used before applying an average pooling layer.

To train and evaluate the models, a portion of the data-set was first labelled manually by three independent human annotators. We selected 1000 (2.7%) randomly chosen tweets (excluding retweets) for the manual labeling procedure and removed 99 tweets in the pre-test phase (these tweets were text-wise duplicated entries that only differed in the screen-

Label	Absolute Frequency	Relative Frequency
1 (Hate Speech)	147	0.17
0 (Non-Hate Speech)	732	0.83

TABLE I
LABELLING OF THE TRAINING DATA.

name of a user mentioned in the tweet). The annotators were given the task to label the remaining 901 randomly selected tweets as either hate speech (1) or non-hate speech (0). We instructed the annotators to label tweets as hate speech if the message:

- is directed towards an individual or a group of individuals,
- aims to seriously disparage or attack the other(s) by abusing the target using profane words, emotionally harming the target or inciting harm and promoting or justifying intolerance or violence against the target, and
- conveys abuse which is related to some protected characteristics (e.g., race, color, ethnicity, gender, sexual orientation, nationality, religion, disability).

Some example tweets that have been labeled as hate speech read: *"All Muslims are terrorists!"*, *"Fu**ing Muslims!"*, *"Refugees cause terrorism! Send them home!"*. Some example tweets that have not been labeled as hate speech read: *"Islamist Terror is our common enemy!"*, *"Terrorism is a blot on humanity."*, *"We need to fight against islamist terrorism!"*[1].

After the labelling procedure was complete, we resolved any remaining annotator discrepancies. For 22 messages, the context was unclear as there was a picture or a video attached to the comment which was not extracted via the Search API. Therefore, these 22 comments have been removed from the data-set. This resulted in a total of 879 comments used for the training data-set. The pair-wise average rater agreement score (Cohen's Kappa) was a decent $\kappa = 0.77$. We applied a majority vote to finally label tweets as either hate speech or non-hate speech (see Table I).

We used the holdout cross-validation method with 80% of data for training and 20% for testing [19], [27]. We applied the three models (lexicon, SVM with a sigmoid kernel, convolutional neural networks (CNN)) on the labelled data and report on the achieved accuracy in Table II.

Re-construction of the direct messaging network. The convolutional neural network (CNN) with random embedding achieved the highest accuracy and was therefore used in our analysis to predict the labels for the remaining data-set. After removing retweets, we re-constructed the direct messaging network to be able to analyze the messaging behavior between senders and receivers. This resulted in two networks – one that contains the senders and receivers of hateful tweets, and one that does not contain hateful tweets.

Model	Features	Accuracy
lexicon	-	0.5654
SVM	BOW with unigrams	0.8239
SVM	BOW with bigrams	0.7898
SVM	BOW with trigrams	0.7216
SVM	BOW with POS	0.7205
SVM	TF-IDF with unigrams	0.8295
SVM	TF-IDF with bigrams	0.8352
SVM	TF-IDF with trigrams	0.8352
SVM	TF-IDF with POS	0.8352
CNN	Random Embedding	0.8693
CNN	GloVe	0.8636

TABLE II
ACCURACY OF HATE SPEECH DETECTION MODELS.

Data analysis and synthesis. To evaluate the important characteristics of hate speech, we analyzed 1) the characteristics of the accounts who actively disseminate hate speech, 2) the characteristics of hateful tweets, 3) the characteristics of communication networks resulting from hate speech, and 4) representative messaging patterns.

IV. RESULTS

As shown in Figure 2, Twitter users predominantly tweeted during the event and its immediate aftermath. The first day of the extraction period represents only 4 hours of the day (the terror attack happened at approx. 8:00pm). Afterwards, the number of related tweets consistently decreases and drops to almost zero by the end of the data extraction period.

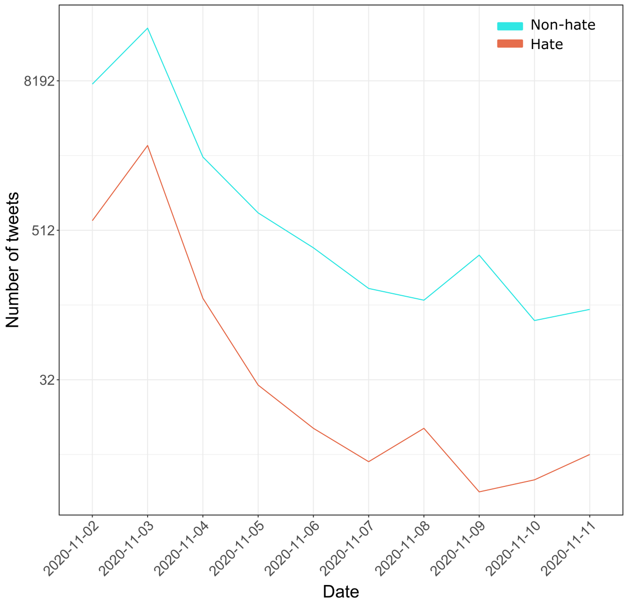

Fig. 2. Number of tweets per day.

With respect to the accounts who spread hateful messages, we evaluated their anonymity by using two characteristics: we checked a) if an account is verified and b) if it includes a link to an external Web page which allows other users to

[1] A note: Special distinction is hereby set on the protected characteristics. While blaming is evident in the example tweets (in this case against Islamic terrorists), the tweet does not contain hate speech. However, those tweets that spread hatred about certain individuals, nations, or religions are considered hate speech. Moreover, we also found that anger and blame were directed at Austrian politicians or the president of Turkey (e.g., "This is Erdogan's fault! He let these terrorists in!!! #boycottturkey"). The annotators did not consider these messages as hate speech because they are not related to some protected characteristic of the person targeted in the tweet.

identify the person behind the account[2]. As shown in Table III, users who spread hate tend to act anonymously on Twitter. However, it must also be noted that most users in our data-set are considered anonymous. Thus, anonymity is not the best indicator for the authorship and dissemination of hate on Twitter.

	Hate Speech	Non-Hate Speech
Verified	0.64%	1.93%
Link to a Webpage	13.98%	21.07%

TABLE III
ANONYMITY OF TWITTER USERS.

With respect to the content of the tweets, we first removed duplicates (i.e., retweets) and stop words from our data-set. We then generated the document term matrix for two separate subsets of our data-set (hateful tweets and non-hateful tweets). We excluded words that appear in both subsets in order to generate a list of unique words that help to distinguish hateful messages from other messages. Figure 3 shows the most common verbs and nouns for each subset normalized by the overall number of messages.

Table IV shows that hateful tweets received slightly more likes than non-hateful tweets on average. However, there is a noticeable difference in the maximum number of likes. In general, our analysis shows that hateful tweets do not reach the number of likes given to non-hateful tweets ($max(n_{likes}(hate)) = 2861$ vs. $max(n_{likes}(nohate) = 4741)$). However, we also confirmed that hateful tweets polarize more than the non-hateful tweets. On average, hateful tweets have a slightly higher number of 0-likes ($n_{0-likes}(hate) = 0.88$ vs. $n_{0-likes}(nohate) = 0.84$) and likes with extreme values ($n_{extreme-likes}(hate) = 0.0012$ vs. $n_{extreme-likes}(nohate) = 0.0010$) (see also Figure 4).

Another trend can be observed when looking at the retweets. On average, hateful tweets result in more retweets than the non-hateful tweets (see Table IV). The higher retweeting rate for hateful content is also seen in Figure 4. We found that 8.44% of the hateful tweets are shared more than 500 times, while the non-hateful tweets showed a comparatively smaller (7.14%) number of tweets that were shared more than 500 times. This indicates that even though hate is only published by a few users ($n_{users}(hate) = 11\%$, $n_{users}(no - hate) = 93\%$, see Table IV), hate can spread widely as it is shared more frequently than the non-hateful tweets.

Next, we derived a direct messaging (DM) network that contains all senders and receivers involved in the exchange of hate (nodes represent Twitter users, edges represent hateful messages)[3]. Basic information about the hate-exchange network and the non-hate exchange network is shown in Table V. Our results indicate that a comparatively small number of users is responsible for the active spread of hate compared to non-hateful tweets.

	Hate Speech	Non-Hate Speech
Likes (mean, stdev)	3.76±71.14	3.45±57.44
Likes (max)	2861	4741
Retweets (mean, stdev)	163.96±243.21	152.24±233.71
Retweets (max)	922	922
Number of users	11%	93%
Tweeting rate	1.34±0.95	1.34±1.45

TABLE IV
USER REACTIONS (LIKES, RETWEETS) TO HATEFUL AND NON-HATEFUL TWEETS.

Figure 5 shows that the number of direct messages fluctuates over time. The first two days of our data extraction period exhibit the highest volume of messages. In the following days the number of direct messages decreases with occasional spikes on certain days (attributed to the availability of new information about the event). Interestingly, we found that on average the volume of hateful directed messages is higher than the non-hateful messages (see Figure 5).

	Hate Speech	Non-Hate Speech
Senders	139 (0.27)	1213 (0.42)
Receivers	376 (0.73)	1710 (0.60)
Edges	563	3309

TABLE V
BASIC INFORMATION ABOUT THE DIRECT MESSAGING NETWORKS INCLUDING THE ABSOLUTE AND RELATIVE NUMBER OF SENDERS AND RECEIVERS, AS WELL AS EDGES (MESSAGES). THE RELATIVE NUMBER OF SENDERS AND RECEIVERS IS AVERAGED OVER THE TOTAL NUMBER OF UNIQUE USERS WHO PARTICIPATE IN A DM NETWORK.

Figure 5 also shows the number of active (senders) and passive (receivers) users in content spreading. In both subsets, there is a higher number of receivers than senders, indicating that the active spreaders address (tag) more users in their direct messages. This behavior also appears to be more prominent with respect to hateful messages, i.e. on average there are more receivers who are tagged in hateful messages compared to non-hateful messages. Moreover, Figure 5 shows that most of the receivers of hateful messages were tagged on the second day (when the perpetrator was identified in the news) and fourth day (when the motivation of the attacker became clearer) after the event.

In order to find patterns that serve as basic building blocks of the daily direct messaging (DM) networks (see Figure 6), we utilized the concept of network motifs [23]. For motif detection, we used the exact enumeration of all possible 3-node subgraphs. Moreover, we used the stub matching algorithm [25] to generate 1,000 null models for each day of the data extraction period and subsequently identified those subgraphs that are statistically significant for our real-world networks. In total, we identified 28 different motifs (distinguished by edge direction and edge weight). Figure 7 shows some simplified examples of the motifs we identified [4].

[2] Note that while the profile picture could be taken into account, we did not consider it in this study since we cannot make valid conclusions about the legitimacy of the photo

[3] The direct messaging network was derived after removing the retweets from our data-set.

[4] In this paper, simplified motifs are motifs that do not include edge weights.

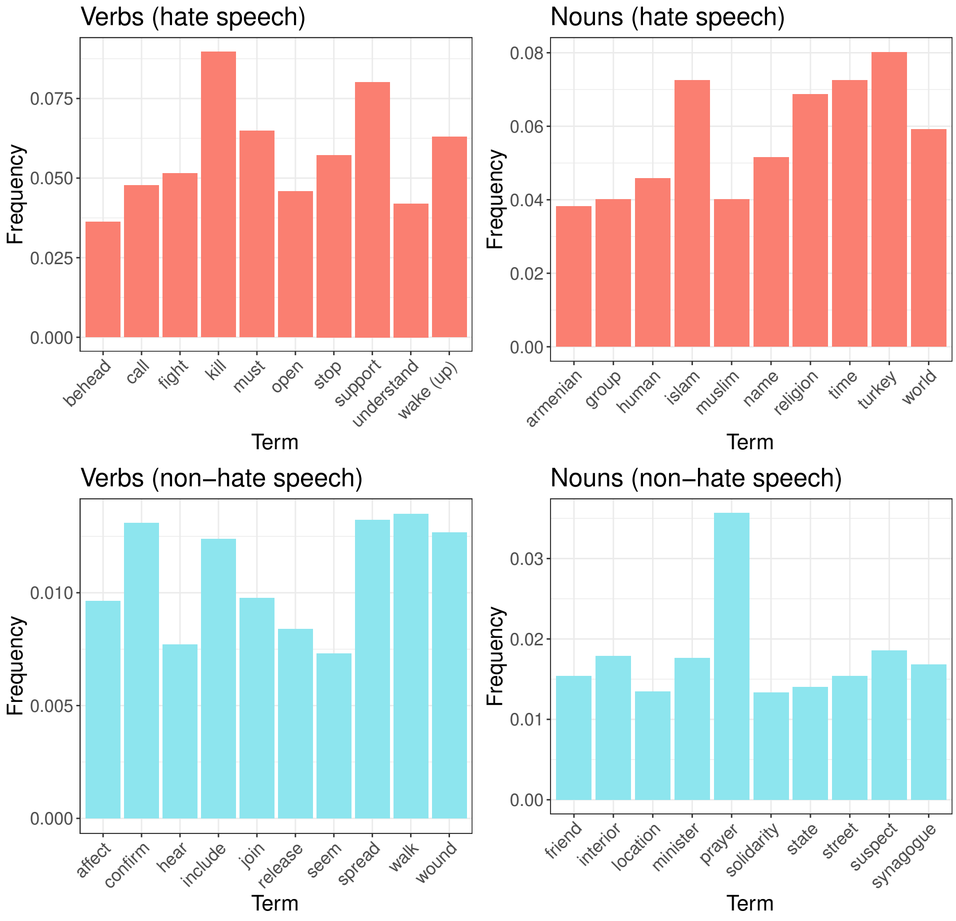

Fig. 3. Common verbs and nouns in hateful and non-hateful tweets (normalized by the total number of messages).

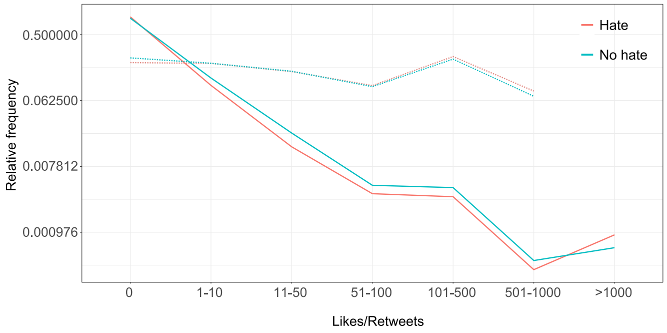

Fig. 4. Average number of favorites (likes) and retweets for hateful and non-hateful tweets. Retweets are shown via a dashed line and likes via a solid line.

We found two motifs that especially represent the exchange of hate, both of which include multiple messages sent from a single user. Moreover, both motifs (in their simplified form) take a common shape presented as the orange graph in Figure 7. Thus, in our data-set a characteristic messaging behavior for hateful messages follows a pattern where one user spreads multiple hateful messages to other users who passively receive the content without any further engagement. In other words, we did not observe bilateral engagement/discussions in network motifs that are characteristic for the spread of hateful messages during the 2020 Vienna terror attack.

Regarding the motifs representing the exchange of non-hateful messages, there is a broad variety of motifs that was detected in our data-set including the bilateral communication between a pair of users, messaging cycles, messaging chains, as well as the presence of self-loops (see the turquoise motifs in Figure 7). Regarding the common simplified motif found in the hate as well as the non-hate network ($A \leftarrow B \rightarrow C$), the average edge weight in the hate-motif is 4.2 while its non-hate counterpart shows the mean edge weight of 5.93.

V. Discussion

When comparing the three approaches for the detection of hate speech that we used in our case study, the lexicon-based approach did not perform well as only 64% of the labels (hate/non-hate) could be accurately predicted. However, during the content analysis of the hateful tweets in our data-set, we observed that there is a number of terms that frequently appear in hateful tweets. This indicates that a more individualized lexicon focusing on terror attacks could be derived to achieve a better accuracy of the lexicon-based approach. However,

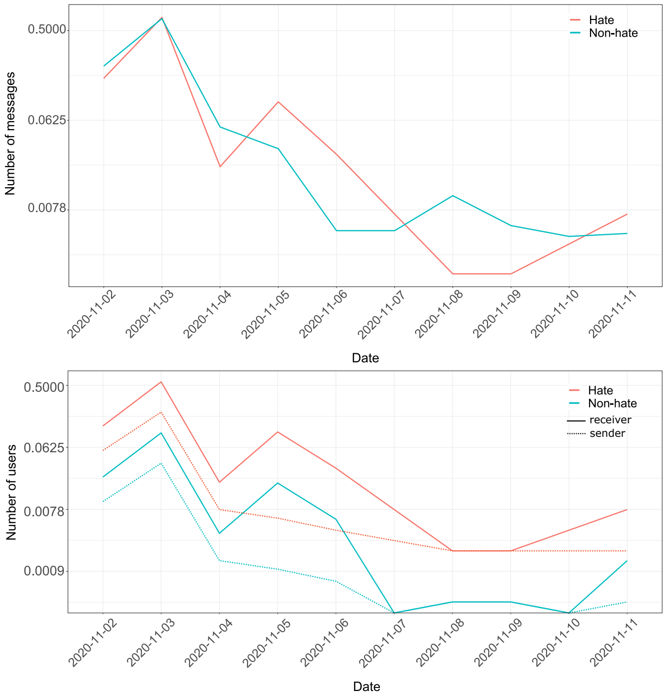

Fig. 5. Temporal flow of the number of direct messages, senders, and receivers of non-hateful and hateful tweets.

it is difficult to identify common words without having a word corpus resulting from various similar events. However, although the lexicon-based approach is easy to understand and easy to implement on a new text corpus, the transferability of a lexicon is questionable as hateful words can differ considerably between different types of events.

In our study, the Support Vector Machine (SVM) approach outperformed the lexicon-based approach. Although the accuracy of the SVM is substantially higher and the model is able to correctly predict many labels, additional features could further improve the accuracy of this model. While sentiment features received only little attention in previous research and were used in a limited number of studies only (see, e.g., [31]), they could potentially improve the accuracy of hate speech detection because hateful tweets are associated with negative sentiments or emotions (such as anger or disgust). In contrast, tweets that express compassion or plead for solidarity are associated with positive sentiments.

Due to the nature of the event that we analyzed, our data-set contains more hate speech than regular everyday tweets. Therefore, the findings presented in this paper cannot be generalized. However, they do comply with findings reported in the related work. In general, the increasing use of multimodal content, such as videos, photos, GIFs, or the large variety of emojis that can make a comment appear racist for example, brings a new challenge for detecting hate speech. For example, [15] considered several components of a tweet and analyzed not only the text, but also the images and the text in the images.

Our findings indicate that accounts who spread hateful messages are more likely to show certain characteristics. We found that less than 1% of the accounts spreading hateful messages are verified Twitter accounts (in other words, more than 99% of the hate spreaders are *not* verified accounts). This confirms the findings of [24] who reported that an increasing degree of anonymity increases the aggressiveness in OSNs. In addition, [34] found that a higher likelihood of posting hateful content is correlated with a higher level of anonymity of the users. This finding is attributed to the phenomenon of *disinhibition* when interacting online. Since much of the content that is published online cannot be easily traced back to the actual identity of the user, the corresponding users are never pressured to explain him- or herself [34].

When considering the user behavior upon encountering hate speech, we found that hateful tweets were retweeted more often than the non-hateful tweets. In general, this shows the potential of hate speech to widely spread across the network. Thereby, we can confirm the findings of [4] who reported that hate speech can spread widely via retweeting.

Our findings further showed that the emergence of hate happens in bursts, i.e. it happens rapidly within a short period of time, decreases in the aftermath of the event, but has the potential to reappear at any time. The temporal patterns that we observed can be explained in the light of the theory of four stages of hate speech dissemination [8]. In particular, we found that a considerable amount of hate speech occurs directly after the event (called the *influence stage* [8]). This could be associated with the high level of affective arousal[5] of those affected by the terror attack.

According to [8], the influence stage is followed by the *intervention stage* where the amount of hate speech decreases and reaches a low level, resulting in the *response stage*. Although there are some days in our data-set where the number of hateful messages slightly increases (November 8 and November 11), the deviation from the decreasing trend is not substantial. As the data for this study was only extracted till 10 days after the attack, the results of the *rebirth stage* [8] cannot be analyzed in this paper. However, it would be interesting to extend this study over a longer time period. For example, [38] analyzed the short-term and the long-term consequences of terror attacks and found that people are anxious and fearful up to some months after the event. Thus, one question that we leave for future work is whether hate will still (re)emerge after a longer period of time and whether its structural patterns and dissemination characteristics will be comparable to the findings in this paper.

A. Limitations

Our paper is based on a data-set related to the 2020 Vienna terror attack and therefore the results cannot be generalized for other types of events that cause the emergence of hate speech. While other terrorist attacks are more likely to show similar effects on Twitter to the ones discovered in this paper, communication patterns emerging in the context of non-terrorist events might look different. However, such aspects

[5] "Affective arousal describes the state of feeling awake, activated, and highly reactive to stimuli." [26]

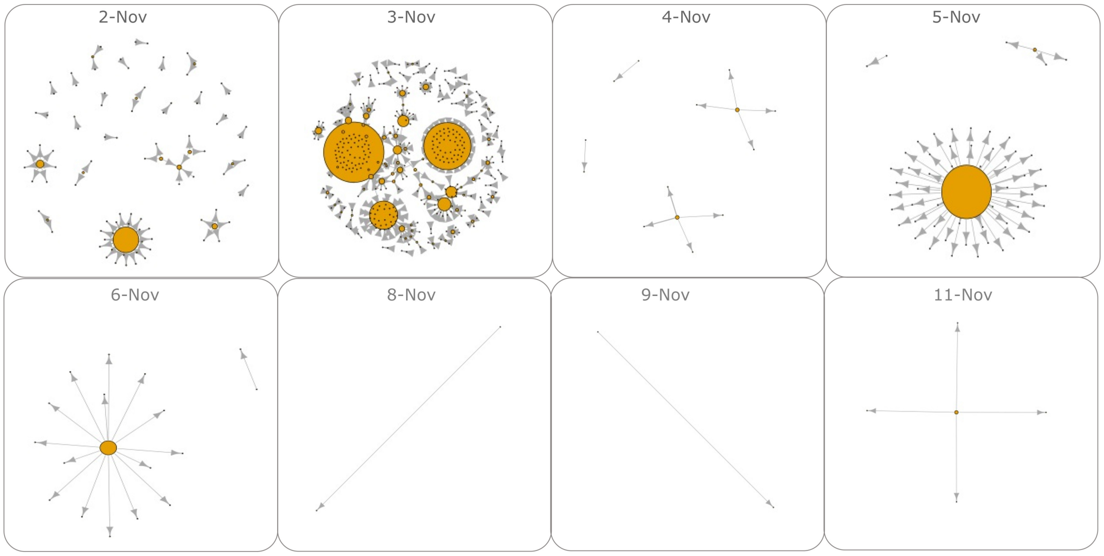

Fig. 6. Daily hateful communication networks.

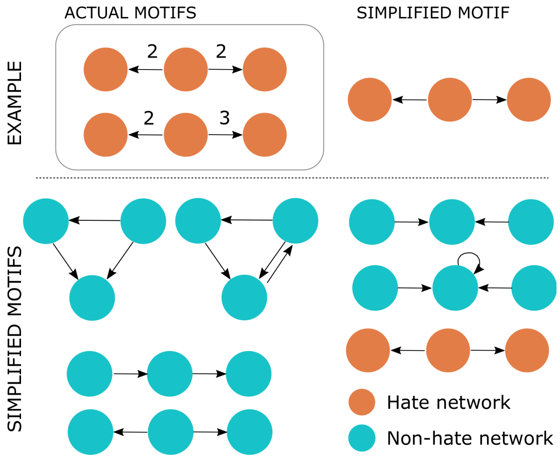

Fig. 7. Motifs detected in hate and non-hate networks.

could not be taken into consideration by analyzing a single case study only.

Another limitation lies in the selection of the key-words and hashtags that we used for data extraction and the rate limits imposed by Twitter. Therefore, we cannot exclude the possibility that we missed some relevant tweets.

Moreover, regarding the motif detection, the exact impact of a different null model generation algorithm on the corresponding subgraph distribution is unclear.

VI. Conclusion

In this paper, we analyzed a data-set consisting of 36,685 tweets that have been sent in the aftermath of the 2020 Vienna terrorist attack. We used several approaches to detect hate speech and found that the CNN approach performed with the highest accuracy. We also found that most tweets directly follow after the triggering event has occurred while the number of related messages rapidly decreases in the days after the event.

We also found that Twitter accounts responsible for the active spread of hateful messages are predominantly anonymous, and that hateful comments are more commonly retweeted compared to other tweets. Our analysis further revealed one significant pattern for the spread of hate (a message-broadcaster pattern). Moreover, we also found that other types of network motifs exist which only emerge in non-hate networks.

In our future work, we aim to work on the improvement of different hate speech detection techniques. In addition, it would be interesting to examine the role of social bots with respect to the spreading of hateful messages.

References

[1] Aymé Arango, Jorge Pérez, and Barbara Poblete. Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems*, 2020.

[2] Nofa Aulia and Indra Budi. Hate speech detection on indonesian long text documents using machine learning approach. In *Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence*, pages 164–169, 2019.

[3] 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.

[4] Pete Burnap and Matthew L Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242, 2015.

[5] Pete Burnap and Matthew L Williams. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data science*, 5(11):1–15, 2016.

[6] Ayon Chakraborty, Jyotirmoy Sundi, Som Satapathy, et al. Spam: a framework for social profile abuse monitoring. *CSE508 report, Stony Brook University, Stony Brook, NY*, 2012.

[7] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*, pages 71–80. IEEE, 2012.

[8] Naganna Chetty and Sreejith Alathur. Hate speech review in the context of online social networks. *Aggression and violent behavior*, 40:108–118, 2018.

[9] Maral Dadvar, FMG de Jong, Roeland Ordelman, and Dolf Trieschnigg. Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*. University of Ghent, 2012.

[10] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*, pages 693–696. Springer, 2013.

[11] Maral Dadvar, Rudolf Berend Trieschnigg, and Franciska MG de Jong. Expert knowledge for automatic detection of bullies in social networks. In *25th Benelux Conference on Artificial Intelligence, BNAIC 2013*, pages 57–64. Delft University of Technology, 2013.

[12] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*, 2017.

[13] Facebook. Controversial, Harmful and Hateful Speech on Facebook. https://m.facebook.com/notes/facebook-safety/controversial-harmful-and-hateful-speech-on-facebook/574430655911054, September 2020.

[14] Robert James Gabriel. Google profanity words. https://github.com/RobertJGabriel/Google-profanity-words/blob/master/list.txt, n.d. Github.

[15] R. Gomez, J. Gibert, L. Gomez, and D. Karatzas. Exploring hate speech detection in multimodal publications. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1459–1467, 2020.

[16] Hatebase. Hate speech. https://hatebase.org/, November 2020.

[17] Alex Hern. Facebook, YouTube, Twitter and Microsoft sign EU hate speech code. https://www.theguardian.com/technology/2016/may/31/facebook-youtube-twitter-microsoft-eu-hate-speech-code, May 2016. The Guardian.

[18] Muhammad Okky Ibrohim, Muhammad Akbar Setiadi, and Indra Budi. Identification of hate speech and abusive language on indonesian twitter using the word2vec, part of speech and emoji features. In *Proceedings of the International Conference on Advanced Information Science and System*, pages 1–5, 2019.

[19] S. T. Luu, H. P. Nguyen, K. Van Nguyen, and N. Luu-Thuy Nguyen. Comparison between traditional machine learning models and neural network models for vietnamese hate speech detection. In *2020 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 1–6, 2020.

[20] Toni M Massaro. Equality and freedom of expression: The hate speech dilemma. *Wm. & Mary L. Rev.*, 32(211):211–265, 1991.

[21] Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*, pages 173–182, 2019.

[22] Microsoft. Report Hate Speech Content Posted to a Microsoft Hosted Consumer Service. https://www.microsoft.com/en-us/concern/hatespeech, October 2020.

[23] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[24] Silvana Neshkovska and Zorica Trajkova. The essentials of hate speech. *Teacher*, 14(1):71–80, 12 2018.

[25] Mark Newman. Networks: An introduction. In *Networks: An Introduction*. Oxford University Press, 2010.

[26] Niven, Karen and Miles, Eleanor. *Affect Arousal*, pages 50–52. Springer New York, New York, NY, 2013.

[27] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153, 2016.

[28] John T Nockleby. Hate speech. *Encyclopedia of the American constitution*, 3(2):1277–1279, 2000.

[29] NoSwearing. List of swear words, bad words, & curse words. https://www.noswearing.com/dictionary, n.d. Swear Words List & Curse Filter.

[30] Committee of Ministers. Recommendation no. r (97) 20 of the committee of ministers to member states on "hate speech". *COUNCIL OF EUROPE*, 1997.

[31] O. Oriola and E. Kotzé. Evaluating machine learning techniques for detecting offensive and hate speech in south african tweets. *IEEE Access*, 8:21496–21509, 2020.

[32] Sergio A Rojas-Galeano. Revealing non-alphabetical guises of spam-trigger vocables. *Dyna*, 80(182):15–24, 2013.

[33] Rstudio. Text Classification. https://tensorflow.rstudio.com/tutorials/beginners/basic-ml/tutorial_basic_text_classification/#build-the-model, 2020.

[34] John Suler. The online disinhibition effect. *Cyberpsychology & behavior*, 7(3):321–326, 2004.

[35] Twitter. Hateful Conduct Policy. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy, March 2020.

[36] Luis von Ahn. Offensive/profane word list. https://www.cs.cmu.edu/~biglou/resources/, n.d. Carnegie Mellon University.

[37] William Warner and Julia Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26, 2012.

[38] Rachel Yehuda and Steven E Hyman. The impact of terrorism on brain, and behavior: what we know and what we need to know. *Neuropsychopharmacology*, 30(10):1773–1780, 2005.

[39] YouTube. Hate speech policy. https://support.google.com/youtube/answer/2801939, September 2020.