THREAT: A Large Annotated Corpus for Detection of Violent Threats

1st Hugo L. Hammer
Department of Computer Science
OsloMet – Oslo Metropolitan University
Oslo, Norway
hugo.hammer@oslomet.no

2nd Michael A. Riegler
Simula Metropolitan Center for Digital Engineering
Oslo, Norway

3rd Lilja Øvrelid
4th Erik Velldal
Department of Informatics
University of Oslo
Oslo, Norway

***Abstract*—Understanding, detecting, moderating and in extreme cases deleting hateful comments in online discussions and social media are well-known challenges. In this paper we present a dataset consisting of a total of around 30 000 sentences from around 10 000 YouTube comments. Each sentence is manually annotated as either being a violent threat or not. Violent threats is the most extreme form of hateful communication and is of particular importance from an online radicalization and national security perspective. This is the first publicly available dataset with such an annotation. The dataset can further be useful to develop automatic moderation tools or may even be useful from a social science perspective for analyzing the characteristics of online threats and how hateful discussions evolve.**

***Index Terms*—national security, publicly available dataset, social media, threat detection, violent threats**

I. Introduction

Internet and social media are giving individuals a unique arena to participate in all sorts of discussions. This has created new opportunities for information and opinion sharing. Unfortunately online discussions are often contaminated by abominable behavior like making violent threats [3], [9], a development which clearly is a cause for concern. Social media providers thus struggle to provide good services to their users and often online threats are directed towards women, kids and vulnerable minorities [8], [18]. Further, the posing of violent threats is illegal under both international and national laws [24] (Article 20). There exist examples of individuals who have been arrested for expressing hate or threats on the Internet [6], [19], [25]. From a national security perspective one may be concerned that hateful online discussions may result in radicalization [1], [21], [22].

Social media providers are trying to alleviate these challenges by using moderators. Moderators either remove adverse comments or send them back to the users to give them the chance to reformulate. However, such tasks are both time consuming and costly. From a national security and law enforcement perspective, it is impossible to manually monitor even a fraction of the enormous amount of activity on social media.

To address these challenges, several papers have suggested tools that automatically detect threats and other abominable commenting [1], [2], [5], [11], [12], [15], [16], [26]. The methods are mainly based on machine learning and thus require annotated text to learn to separate abominable from harmless online behaviour. Unfortunately, neither of these studies have made the accompanied datasets publicly available. In fact, we are not aware of any publicly available datasets that can be used to develop automatic threat detection.

As a contribution to solve these challenges and to make it possible to perform open and important research on making cyberspace more secure for people we present a large dataset of YouTube comments, where each sentence (manually segmented) is annotated as either being a threat of violence or not.

Previous and non-publicly available versions of the dataset presented in this paper have been analyzed in [11], [12], [20], [26] for the task of threat detection. However, the dataset now made publicly available contains significant improvements and extensions compared to the non-publicly available datasets previously analysed. First, and most importantly, each annotated sentence is now associated with two new attributes, namely a timestamp when the comment was published and to which video the sentence belonged to. The new attributes open up for new and interesting research directions that will be further described below. Second, the dataset is now completely anonymised and, third, some sentences that were erroneously annotated as non-threats in the original version have been corrected.

In the non-publicly available datasets previously analyzed each annotated sentence were associated with comment id and user id (anonymised). We now give examples on how the new attributes (associated video and timestamp) opens up new and interesting research directions.

A well-known problem with machine-learned classifiers, for related tasks like sentiment analysis, is that performance can be expected to drop, sometimes substantially, when applying the classifier to material that is from another domain than the annotated training data. Problems posed by out-of-domain applications is typically made worse if the training dataset is small and non-diverse. The dataset in this paper consists of comments from 19 different videos related to quite diverse religious and political topics like e.g. the Arab-Israeli conflict, the Eurabia theory [7] or halal slaughter. By training a classi-

978-1-7281-4673-7/19/$31.00 ©2019 IEEE

fier on comments from some videos to detect threats on others gives an opportunity to analyze and develop classifiers with the focus on robustness against domain changes.

The new timestamp attribute makes it possible to study how threats progresses in online discussions. This can possibly be used to develop tools for early detection of escalating aggression in discussion fora.

It is important to state that any development of automatic moderation or monitoring tools must be weighed against individuals rights to freedom of expression [9].

The main contributions of the paper and the accompanying dataset are:

1) Presentation and sharing of a large scale dataset of online YouTube video comments where each sentence is annotated as either being a violent threat or not.
2) Each sentence is associated with a specific comment, user, video and timestamp which opens up for many interesting research directions.
3) The attributes video and timestamp are new and never used in previous research.
4) Our previous research on automatic threat detection based on previous versions of the dataset forms a natural benchmark for future research into automatic threat detection [26].

II. Related work

In Sections II-A, we present related research and in Section II-B, we present related datasets.

A. Related research

To the best of our knowledge, the only previously reported study based on YouTube comments is that of Dinakar (2011) [5]. 4500 comments from YouTube videos involving what was deemed sensitive topics related to race & culture, sexuality and intelligence were annotated to indicate whether they could be seen as negative remarks along those same dimensions (e.g. negative comments towards sexual minorities or women). They then report results for trying to detect 'cyberbullying' on the basis of this, but unfortunately the dataset itself was not made publicly available.

While there appears to not be much previous work specifically targeting detection of violent threats, it is a growing interest for research on hate-speech and online harassment, typically with a focus on social media. Among the shared tasks of the International Workshop on Semantic Evaluation this year (SemEval 2019) we find detection of phenomena like offensive language, hate speech and hyper-partisan argumentation in news. Another important forum for research on related topics is provided by the Workshop on Abusive Language Online[1] (ALW) which will be held for the third time in 2019.

[1]https://sites.google.com/view/alw2018/

B. Related publicly available datasets

First we discuss some *comment*-based corpora that are available and contain annotations of related phenomena like hate-speech and harassment. However, please note that our dataset separates from all these datasets by focusing on violent threats which is the most extreme form of hate-speech and is particularly important from an online radicalization and national security perspective.

Examples of relevant datasets include the various annotations added to parts of the Wikipedia Comments Corpus, prepared by the Wikipedia Detox project.[2] This comprises over 100K comments annotated for personal attacks, aggression, and toxicity (with 10 crowd-sourced judgments per comment) [28].

Another dataset of annotated comments is the Yahoo News Annotated Comments Corpus (YNACC), in which 9.2k comments and 2.4k threads posted in response to Yahoo News articles have been annotated in terms of several dimensions including constructiveness, agreement, tone, sentiment, type (e.g. Argumentative, Flamewar, Positive/respectful, Off-topic, etc.) and more [14].

The SFU Opinion and Comments Corpus[3] (SOCC) is a corpus of more than 300K threads comprising over 660K comments posted to opinion articles (editorials, columns, and op-eds) that have been annotated for four different phenomena: constructiveness, toxicity, negation and its scope, and appraisal [13].

There is also other work directed towards detection of hate-speech that is not based on comment data, like the corpus of tweets annotated for hate speech and offensive language[4] provided by [4]. There are also several other Twitter-based datasets, for instance for detection of trolling [23] or harassment [10] or identification of hateful users [17].

III. Dataset collection and details

The dataset consists of comments from 19 different YouTube videos.[5] Each video was related to religious and political topics that typically created a lot of anger and disagreements like the Arab-Israeli conflict, Eurabia theory [7], halal slaughter, Anders Behring Breivik, Geert Wilders etc. The dataset was collected during the summer of 2013.

Each sentence in the material was manually annotated to either contain a threat of violence (or sympathy with violence) or not. For sentences where it was impossible to decide, e.g., due to terribly poor language, or the sentence was part of a larger argument, the sentence were annotated as a non-threat. A few comments contained copies of violent passages from the Bible or the Quoran. Such sentences were classified as violent if the passage was violent.

[2]https://meta.wikimedia.org/wiki/Research:Detox/Data_Release

[3]https://github.com/sfu-discourse-lab/SOCC

[4]https://data.world/crowdflower/hate-speech-identification

[5]In [11] (and further referred to in [12], [26]) it was stated that the comments came from eight different videos. However a recent recounting found that this was erroneous and that the correct number of videos is 19

	Comments	Sentences	Users
Total	9 845	28 643	5 484
Threats	1 287	1 387	993

TABLE I
THE NUMBER OF COMMENTS, SENTENCES AND UNIQUE USERS IN THE YOUTUBE THREAT DATASET.

Posted	No. of users
26 threats	1
16 threats	1
12 threats	1
10 threats	1
9 threats	4
8 threats	1
7 threats	7
6 threats	4
5 threats	9
4 threats	13
3 threats	25
2 threats	108
1 threats	818
0 threats	4 491

TABLE II
NUMBER OF THREATS OF VIOLENCE (SENTENCES) POSTED BY USERS.

The text material consists of a total of 28 643 sentences where 1 387 were annotated as being violent threats. Each sentence is part of a comment, posted by a user, at a specific time to a specific video and each sentence is associated with these attributes. The dataset consists of a total of 9 845 comments from 5 484 different users. Please see Tables I and II for further details about the dataset.

Each user was anonymized as follows. Each comment started with the user name which was changed to `Commenter #1`, `Commenter #2`, and so on. There were further references to user in the comment text and if a reference was e.g. made to `Commenter #2`, the user name was changed to `@Commenter#2`. There were even some references to users that had not made comments in the dataset. Since the dataset consisted of comments from 5 484 users, these references were changed to `@Commenter#5485`, `@Commenter#5486`, and so on.

A randomly selected subset of 100 non-violent and 20 violent sentences (based on the annotation from the main annotator) were labeled by another annotator for inter-annotator studies. The results showed that for 98% of the sentences, both annotators made the same decision, and both annotators found a total of 20 threats of violence. For further details about the dataset, we refer to [26], [27].

A short example taken from the dataset is shown in Figure 1. Sentences starting with '`1`' were annotated as being a threat (or sympathy with violence) while other sentences were annotated with '`0`'.

To download the dataset, please fill out the following form https://docs.google.com/forms/d/e/1FAIpQLScQTVDqROxIg4YSq1xJHkCkolhXStPbeW3gricJprNkTQZccw/viewform?usp=sf_link where you agree that you will only use the dataset for academic purposes and that you will delete the dataset on request. When you submit the form we will provide you with a download link.

```
Video #1, Comment #93, User #52, 2 months ago
1 and i will kill evey fucking muslim and arab!

Video #1, Comment #94, User #51, 2 months ago
0 You have fun with that, pal.
0 The world is, overall, moving in a completely
  different moral direction than that.
0 Have fun hating on another race in the comforts
  of the community of this video.
```

Fig. 1. A short section from the YouTube threat dataset. Please note the writing errors were not corrected and are as collected.

IV. APPLICATIONS OF THE DATASET

Our vision is that the available dataset can help to develop improved tools to alleviate the substantial challenges with hate speech and threats in social media and online discussions.

As pointed out previously we believe that the dataset can open up many interesting research directions. To the best of our knowledge, this is the first annotated dataset focusing on violent threats. Violent threats is the most extreme form of hateful communication and we believe the dataset can be particularly useful from perspectives of online radicalization and national security. We also believe the dataset can be used to develop new and improved threat detection and automatic discussion moderation tools. Classifiers can be trained and evaluated on at least three different levels; the sentence-, comment- or user-level. In terms of the latter, the material can also be used for automatic risk assessments of users in terms of radicalization and national security. Furthermore, since the material is divided into different videos we believe the dataset is unique to develop models that are robust to domain changes. Finally, a completely different approach could be to use the dataset to study differences in the writing style or quality of users making hateful comments compared to others.

V. SUGGESTED METRICS

As pointed out in the introduction and in Section IV, the dataset opens up many interesting research directions and questions to answer. Straightforward and natural tasks include detecting threatening users or detecting comments or sentences containing violent threats. All of these tasks can be approached as standard binary text classification tasks, and would correspondingly be evaluated using standard metrics like precision, recall, and F1 score. Note that a metric like accuracy would typically not be well-suited here given the skewed class-distribution. This class imbalance also means that techniques like (over-/under-)sampling, cost-sensitive learning or class-weighted loss functions could be useful during training.

VI. BASELINE PERFORMANCE

Reproducibility and comparability of results is an important factor of high quality research. As pointed out in Section II, earlier versions of the dataset have been analyzed in [11], [12], [20], [26]. In particular [26] performed systematic analysis for the potential of using the dataset for efficient threat detection.

The work of [26], [27] reported on experiments with a range of classifiers based on Maximum Entropy (MaxEnt), Support Vector Machines (SVM), and Random Forest (RF) with a wide range of different features; simple lexical features like bag-of-words and n-grams defined over both full-forms and lemmas, but also morphosyntactic features like PoS-tags and syntactic relations extracted from dependency graphs, and finally class-based features extracted from WordNet synsets and Brown clusters. The experiments showed that simple feature configurations based on only lexical information gave the best performance, with the SVM model yielding an F1-score of 68.85 in held-out testing on the sentence-level [26] (training and evaluation on the comment-level is also possible).

[20] reports results for training various convolutional neural network (CNN) classifiers using the same version of the dataset as [26]. To mitigate the problem of unbalanced classes, the experiments of [20] included scaling of the loss function as to give more weight to the threat-class. Despite large-scale tuning of parameters like the number of filter maps and their window sizes, drop-out, pre-trained word embeddings, and more, the best CNN achieved an F-score of 65.29 [20], thus failing to outperform the SVM model of [26].

The studies discussed above form a natural baseline for future research. The version of the dataset used were not anonymised and therefore is not publicly available, but details about the results and evaluation that allow reproducibility can be obtained by contacting the authors of [26].

VII. Conclusion

In this paper we present a large dataset consisting of comments from 19 different YouTube videos. The videos were about controversial political and religious themes which created a lot of aggressive discussions. The dataset consists of a total of about 30 000 sentences extracted from circa 10 000 comments. For each of the sentences we provide annotations that indicate if the sentence is a violent threat (or sympathy with such) or not. To the best of our knowledge this is the first publicly available dataset of this kind. We believe that the dataset can bring the research on threat detection and online radicalization to new levels of quality. Finally, the authors of this paper point out that any moderation and monitoring of online behaviour must always be weighed up against individuals rights to freedom of expression.

References

[1] Swati Agarwal and Ashish Sureka. Using common-sense knowledge-base for detecting word obfuscation in adversarial communication. In *Communication Systems and Networks (COMSNETS), 2015 7th International Conference on*, pages 1–6. IEEE, 2015.

[2] Swati Agarwal and Ashish Sureka. Using KNN and SVM based one-class classifier for detecting online radicalization on twitter. In *International Conference on Distributed Computing and Internet Technology*, pages 431–442. Springer, 2015.

[3] Naganna Chetty and Sreejith Alathur. Hate speech review in the context of online social networks. *Aggression and violent behavior*, 2018.

[4] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International Conference on Web and Social Media (ICWSM)*, 2017.

[5] Karthik Dinakar, Roi Reichart, and Henry Lieberman. Modeling the detection of textual cyberbullying. *The Social Mobile Web*, 11(02):11–17, 2011.

[6] Euronews. Neo-Nazi and black metal star Varg Vikernes arrested in France. http://www.euronews.com/2013/07/16/neo-nazi-and-black-metal-star-varg-vikernes-arrested-in-france-/, 2013. [Online; accessed 21-January-2018].

[7] Liz Fekete. The Muslim conspiracy theory and the Oslo massacre. Technical Report 53(3): 30 – 47, Institute of Race Relations, 2011.

[8] Jerry Finn. A survey of online harassment at a university campus. *Journal of Interpersonal violence*, 19(4):468–483, 2004.

[9] Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. *Countering online hate speech*. Unesco Publishing, 2015.

[10] Jennifer Golbeck, Zahra Ashktorab, Rashad O Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, et al. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 229–233. ACM, 2017.

[11] Hugo Lewi Hammer. Detecting threats of violence in online discussions using bigrams of important words. In *Intelligence and Security Informatics Conference (JISIC)*, pages 319–319, 2014.

[12] Hugo Lewi Hammer. Automatic detection of hateful comments in online discussion. In *International Conference on Industrial Networks and Intelligent Systems*, pages 164–173. Springer, 2016.

[13] V. Kolhatkar and M. Taboada. Constructive language in news comments. In *Proceedings of the 1st Abusive Language Online Workshop, 55th Annual Meeting of the Association for Computational Linguistics*, pages 11–17, Vancouver, 2017.

[14] Courtney Napoles, Joel Tetreault, Enrica Rosata, Brian Provenzale, and Aasish Pappu. Finding good conversations online: The yahoo news annotated comments corpus. In *Proceedings of The 11th Linguistic Annotation Workshop*, pages 13–23, Valencia, Spain, 2017.

[15] Nelleke Oostdijk and Hans van Halteren. N-gram-based recognition of threatening tweets. In *Computational Linguistics and Intelligent Text Processing*, pages 183–196. Springer, 2013.

[16] Nelleke Oostdijk and Hans van Halteren. Shallow parsing for recognizing threats in dutch tweets. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1034–1041. ACM, 2013.

[17] Manoel Horta Ribeiro. Hateful users on twitter. https://www.kaggle.com/manoelribeiro/hateful-users-on-twitter, 2018. [Online; accessed 22-January-2018].

[18] Anirban Sengupta and Anoshua Chaudhuri. Are social networking sites a source of online harassment for teens? evidence from survey data. *Children and Youth Services Review*, 33(2):284–290, 2011.

[19] Oda Leraan Skjetne and Halldor Hustadnes. Ubaydullah Hussain is accused of violent threats (norwegian). http://www.dagbladet.no/2013/11/20/nyheter/ubaydullah_hussain/innenriks/islamisme/trusler/30429704/, 2014. [Online; accessed 21-January-2018].

[20] Camilla Emina Stenberg. Threat detection in online discussion using convolutional neural networks. Master's thesis, University of Oslo, 2017.

[21] Øyvind Strømmen. *The Dark Net. On Right-Wing Extremism, Counter-Jihadism and Terror in Europe*. Cappelen Damm, Oslo, Norway, 2012.

[22] Inger Marie Sunde. Preventing radicalization and violent extremism on the Internet (Norwegian). The Norwegian Police University College 2013:1, 2013.

[23] Paraskevas Tsantarliotis, Evaggelia Pitoura, and Panayiotis Tsaparas. Troll vulnerability in online social networks. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1394–1396. IEEE Press, 2016.

[24] UnitedNations. International Covenant on Civil and Political Rights. http://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx, 2014. [Online; accessed 21-January-2018].

[25] Ole Valaker and Magnus Aamo Holte. Bergen blogger arrested (norwegian). http://www.bt.no/nyheter/lokalt/Bergens-blogger-pagrepet-2732162.html, 2012. [Online; accessed 21-January-2018].

[26] Aksel Wester, Lilja Øvrelid, Erik Velldal, and Hugo Lewi Hammer. Threat detection in online discussions. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 66–71, 2016.

[27] Aksel Ladegård Wester. Detecting threats of violence in online discussions. Master's thesis, University of Oslo, 2016.

[28] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*, pages 1391–1399, Perth, Australia, 2017.