Individual vs. Group Violent Threats Classification in Online Discussions

Noman Ashraf
nomanashraf@sagitario.cic.ipn.mx
CIC, Instituto Politécnico Nacional
Mexico City, Mexico

Rabia Mustafa
rabiamustafa954@gmail.com
Independent researcher

Grigori Sidorov*
sidorov@cic.ipn.mx
CIC, Instituto Politécnico Nacional
Mexico City, Mexico

Alexander Gelbukh*
gelbukh@gelbukh.com
CIC, Instituto Politécnico Nacional
Mexico City, Mexico

ABSTRACT

Violent threat is a serious crime affecting the targeted individuals or groups. It is essential for media providers to block the users that post such threats. In this paper, we focused on detection of violent threat language in YouTube comments. We categorized the threatening comments into those targeting an individual or a group. We started from an existing dataset with violent threat language identified, but without any categorization into comments targeting individuals or groups. We adopted a binary classification approach for the prediction of individual- vs. group-targeting threats. We compared two text representations: bag of words (BOW) and pre-trained word embedding such as GloVe and fastText. We used deep-learning classifiers such as 1D-CNN, LSTM, and bidirectional LSTM (BiLSTM). GloVe embedding showed the worst results, fastText performed much better, and BiLSTM on BOW with term frequency-inverse document frequency (TF-IDF) weighting scheme gave the best results, achieving 0.94% ROC-AUC and Macro-$F1$ score of 0.85%.

KEYWORDS

Violent threat, individual and group threats, deep learning, social media, NLP

ACM Reference Format:
Noman Ashraf, Rabia Mustafa, Grigori Sidorov, and Alexander Gelbukh. 2020. Individual vs. Group Violent Threats Classification in Online Discussions. In *Companion Proceedings of the Web Conference 2020 (WWW '20 Companion), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3366424.3385778

1 INTRODUCTION

In today's world, social media is giving abundant opportunity to the people around the globe to share their point of views on every aspect of life. Sharing healthy and proper views on several issues of everyday life is quite a better way to relieve mental and emotional thoughts of people. Unfortunately, some people use this platform to express their violent behavior towards others which could lead some serious actions or to somewhat jeopardize peace of the society [Gagliardone et al. 2015]. That is why cybercrime is becoming serious issue of now a days. It is like a war which is being declared on internet grounds. Article 25 of international law considers that posting threat on internet is a crime [Keith 1999]. Therefore, it has become a matter of concern for the researchers to point out such violent posts on social media so that the relevant authorities may take action against such people.

It is a challenge for social media providers to make it a threat free platform. That is why they need moderators to remove such threats from social media, automatically. Although plenty of moderators based on machine learning are working on several projects, yet this task is laborious and lengthy. Their purpose is to detect those texts which contain threats, hate speech, abusive language, etc. while separating them with harmless data [Agarwal and Sureka 2015]. In order to detect threats from the dataset, there is a need to create dataset or to have the availability of the dataset but it is a hard task because in previous studies dataset is not publically available except [Hammer et al. 2019].

The main contributions of this research are as follows:

- Formulation of the task of classification of violent threats into those targeting an individual or a group.
- Augmentation of a well-known violent threat dataset with such annotation.
- Comparison of deep-learning algorithms (1D-CNN, LSTM, BiLSTM) for this task using various text representations, such as GloVe, fastText, and BOW with TF-IDF.

2 RELATED WORK

Detection of threat is a very crucial task as it has the potential to cause harm to the people concerned. Therefore, a few research studies have been conducted to address this issue. There are a few comment based corpora that comprise annotation of several topics such as threat, abusive, hate-speech etc. A corpus namely The SFU Opinion and Comments Corpus (SOCC)[1] that comprises more than 300K threads containing over 660K comments which were taken from opinion based articles such as columns, editorials etc. Four different phenomena as toxicity, appraisal, constructiveness, and negation

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

WWW '20 Companion, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7024-0/20/04.
https://doi.org/10.1145/3366424.3385778

and its scope have been annotated on this corpus [Kolhatkar and Taboada 2017]. Further, relevant corpus such as Wikipedia Comments Corpus[2] arranged by Wikipedia Detox project contains more than 100K comments. This corpus was annotated for aggression, toxicity, and personal attack with 10 crowd-sourced judgments each comment [Wulczyn et al. 2017].

Another dataset, Yahoo News Annotated Comments Corpus (YNACC) comprises 2.4k threads and 9.2k comments which were posted as reactions to the articles on Yahoo News. This dataset has been annotated in various ways such as agreement, sentiment, tone, type, constructiveness etc [Napoles et al. 2017]. There have been a few studies in the past to detect such language containing offensive, threat, cyber-bullying, hate, insult etc that put the society at the risk of disturbing peace. A research study worked on German tweets with the idea of detecting non-offensive tweets with offensive ones. Furthermore, offensive tweets were sub categorized as insult, profane or abusive [Wiegand et al. 2018].

A few studies can also be seen that worked on hate-speech, threat or violent detection. Bashir Farhan and Mustafa [2019] extracted aggressive behaviour from Twitter data and that aggressiveness might be caused by uncertain life decisions of people.

In the Dutch dataset, the detection of threats has been focused that contained 5k tweets of such kind. Additionally, random tweets were collected in a large number for the purpose of testing and development. Manually constructed detection patterns were utilized in the form of n-gram but detail is not provided which was used to construct such patterns [Oostdijk and van Halteren 2013a,b]. In [Oostdijk and van Halteren 2013b], a manually constructed shallow parser was attached to the system. In such way, the findings were improved to a recall of 0.59 and the precision of 0.39. Cyberbullying was also detected where the combinations of negative and profane words and other pre-decided sensitive topics were targeted. The data was consisted of almost 50k comments from YouTube videos related to controversial ideas. The accuracy was reported from 0.63 to 0.80 but recall or precision was not reported [Dinakar et al. 2011].

A strategy was presented to detect hate-speech from web-text which was user-generated. It depends on machine learning while having combination of template-based features. A word-sense disambiguation task was approached as the same word might be utilized in the context of non-hateful and hateful. In this system, uni, bi, and trigrams features, brown clusters and part-of-speech-tags were used. With a recall of 0.60 and precision of 0.67, unigram features presented the best results. It was suggested that deeper parsing could identify important phrase patterns [Warner and Hirschberg 2012].

3 EXPERIMENTS

3.1 Dataset and Experimental Settings

We used publicly available dataset developed by Hammer et al. [2019]. This dataset has 28,643 sentences and 9,845 comments which were collected from 19 Youtube videos. These videos are related to religion and politics that created much hatred among people. Out of 28K sentences, 1387 comments were labeled as violent threats (or sympathy with violence) while remaining marked as non-threat. Violent passages from Bible and Quran were also classified as threats [Hammer et al. 2019]. According to Wester et al. [2016], this dataset can be evaluated on three different degrees: user-level analysis, comments, and sentences.

As mentioned above, we took violent threat sentences from existing dataset and further classified the sentences into two ways; threatening comments to the individual and to the group. Sample sentences are shown in Table 2. After that, we started annotation process where we got three annotators who labeled the dataset in these two categories. For the purpose of adding sentences in the study according to the required definition, we used majority voting scheme. We did not add those sentences in which two annotators gave disagreement sign. We removed some of the sentences due to duplication. The dataset is publicaly available[3]. The Table 1 represents statistics of the dataset. We treated violent threat detection as a problem of binary classification. This dataset is valuable for researchers in this area; bigger dataset will be more helpful for deep-learning models. We emphasized on deep-learning classifiers and experimented with various kinds of representations, namely, bag of words and word embedding.

Table 1: Dataset statistics

Class	Sentences	Words	Avg. Words
IND	949	15293	4.44
GRP	343	4543	4.29

3.1.1 Weighting scheme. To represent documents mathematically and to extract the most relevant terms of documents Ramos [2003], term frequency inverse document frequency is used. We call such scheme as a weighting scheme. A weighting scheme is used to find out the values for different features and their scaling across features.

TF-IDF weight of term i in document j in a corpus of N documents is calculated as:

$$Weight_{ij} = tf_{ij} \times log\left(\frac{N}{df_i}\right),$$

where (tf_{ij}) is a number of times term i appear in document j and (df_i) is a number of document containing term i.

3.1.2 Features Extraction. For the embedding base features, GloVe[4] and fastText[5] pre-trained models to obtain the input were used. GloVe pre-trained model was trained on a very large corpus of 2B tweets. Using these models, we extracted fixed-length vectors of 300 dimension. If the word was not found, its embedding obtained from random values between $[-0.1, 0.1]$ and bag of words features were also in our consideration. Later, these vectors were used for the training of classifiers.

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

[3]https://github.com/Noman712/violent_threat_detection/blob/master/violent_threats_dataset_github.csv

[4]https://nlp.stanford.edu/projects/glove/

[5]https://github.com/facebookresearch/fastText

Table 2: Sample sentences of individual and group threats

Sentences	Class
hy motherfuckers what r u saying bitch if i get a small chances i will kill that bitch such a big white ass	IND
i wish i could kill this bitch she only shit abt islam with no doubt there will be death treat on her think bf u talk shit ugly bitch	IND
and i will kill evey fucking muslim and arab	GRP
the world need more hitler for kill those muslims	GRP

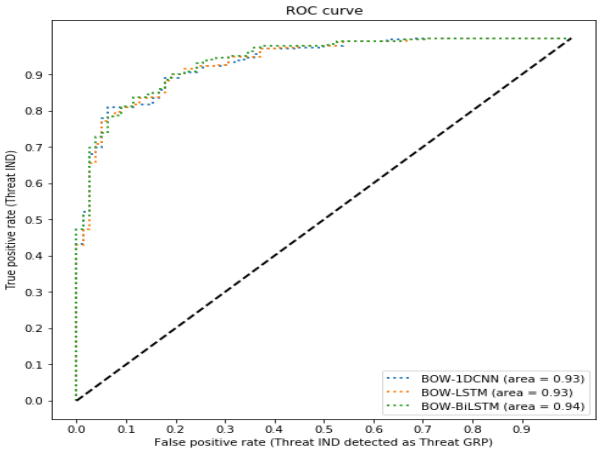

Figure 1: ROC curves using BOW + TF-IDF features on 1D-CNN, LSTM, BiLST

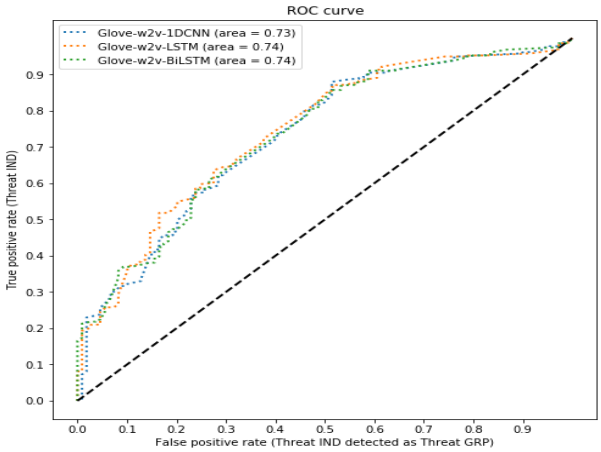

Figure 3: ROC curves using GloVe features on 1D-CNN, LSTM, BiLST

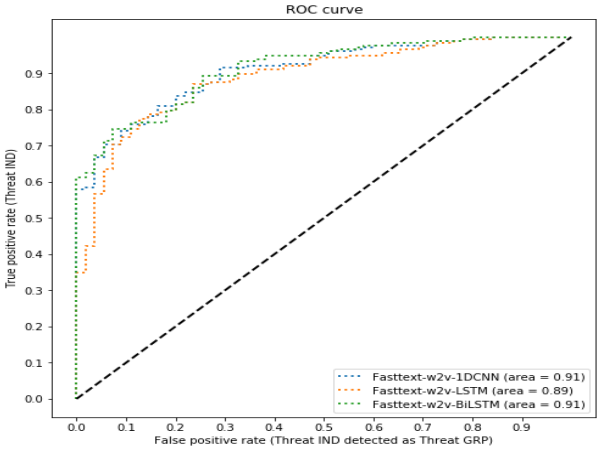

Figure 2: ROC curves using fastText features on 1D-CNN, LSTM, BiLSTM

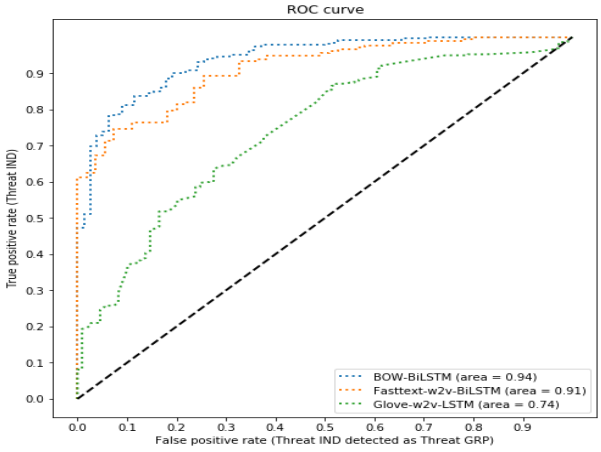

Figure 4: ROC curves comparison between BOW and features on 1D-CNN, LSTM, BiLST

3.2 Setup and Classifiers

The study applied train-test split technique from Scikit-Learn[6] to divide the dataset into train and test samples. From total dataset, 80% was utilized for training and the rest 20% was used for test.

We used three deep-learning classifiers 1D Convolutional Neural Network (1D-CNN), Long short-term memory (LSTM), and Bidirectional Long short-term memory (BiLSTM) for experiments to find the best performing classifier on the dataset. Kim [2014] used CNN for sentiment classification. LSTM networks are a particular kind of RNN that was presented by Hochreiter and Schmidhuber [1997]. They add extra interactions per module to deal with the

[6]https://scikit-learn.org/stable/

Table 3: Results for violent threat language detection. Precision (P), Recall (R), and $F1$-measure ($F1$) for each model on all classes (IND, GRP) are reported. We also listed Macro-$F1$.

		Individual (IND)			Group (GRP)			Weighted Average				
Models	Features	P	R	$F1$	P	R	$F1$	P	R	$F1$	Macro-$F1$	AUC
CNN-1D	BOW + TF-IDF	0.73	0.72	0.72	0.92	0.92	0.92	0.88	0.88	0.88	0.82	0.93
LSTM	BOW + TF-IDF	0.73	0.76	0.74	0.93	0.92	0.92	0.88	0.88	0.88	0.83	0.93
BiLSTM	BOW + TF-IDF	0.77	0.74	0.76	0.93	0.94	0.93	0.89	0.89	0.89	0.85	0.94
CNN-1D	GloVe	0.55	0.42	0.48	0.84	0.90	0.86	0.77	0.79	0.77	0.67	0.73
LSTM	GloVe	0.57	0.39	0.47	0.83	0.91	0.87	0.77	0.79	0.78	0.67	0.74
BiLSTM	GloVe	0.52	0.41	0.46	0.83	0.89	0.86	0.76	0.77	0.76	0.77	0.74
CNN-1D	fastText	0.72	0.53	0.61	0.87	0.94	0.90	0.83	0.84	0.83	0.76	0.91
LSTM	fastText	0.71	0.53	0.60	0.86	0.93	0.90	0.83	0.84	0.83	0.75	0.89
BiLSTM	fastText	0.77	0.62	0.69	0.89	0.94	0.92	0.86	0.87	0.80	0.80	0.91

short comings of RNN [Hochreiter 1998]. LSTM default behavior is to remember information for an extended period as well as long-term dependencies [Le et al. 2019]. LSTM networks are more favorable to the textual data, where the closeness of words might not always be a good benchmark for a trainable pattern. Keras[7] used for the implementation of 1D-CNN, LSTM, and BiLSTM. We used 'Adam' optimizer and 'mean square error' as a loss function for all of our deep-learning classifiers. For additional details on the experiments please review the publicly available code.[8]

3.3 Metrics and Evaluation

For violent threat detection, models performance using Recall (R), Precision (P), and $F1$-measure ($F1$) were evaluated. The mathematical equations of these measures are as follows:

$$Precision = \frac{TP}{TP + FP},$$

$$Recall = \frac{TP}{TP + FN},$$

$$F1\text{-}measure = \frac{2 \times P \times R}{P + R}.$$

4 RESULT ANALYSIS

In this section, experimental results for violent threat detection to the individuals or groups are discussed. We recommended the best classifiers and features. We used deep networks 1D-CNN, LSTM, and BiLSTM for binary classification task. BiLSTM achieved ROC-AUC of 0.94% and Macro-$F1$ score of 0.85%. The results can be seen in Table 3.

In our experiments, the outcomes showed that bag of words performed the best for threat detection classification. GloVe embedding features provided worst outcomes as compared to fastText and bag of words. ROC-AUC curves are shown in the Figures 1 to 4. This might happen due to less amount of training sentences used for the GloVe word embedding. Moreover, violent threats keywords may not be found in the text corpus that used for GloVe training. Another reason that pre-trained models does not work well might be due to they are trained on Twitter. So, transfer-learning techniques might be needed which may improve the results on the tasks related to threat detection.

[7]https://keras.io

[8]https://github.com/Noman712/violent_threat_detection/tree/master/code

5 CONCLUSION AND FUTURE WORK

In this research study, we explored deep-learning for the detection of violent threats to the individuals and groups on YouTube sentences. Our dataset based on the existing threat dataset into individual and group threats were annotated. Further, we investigated two text representations: bag of words (BOW) and pre-trained word embedding and found that deep-learning perform best on BOW with TF-IDF features. Our study achieved 0.94% ROC-AUC and Macro-$F1$ score of 0.85% on BiLSTM. In future, we have a plan to apply context based embedding on comment-level and user-level data.

ACKNOWLEDGMENTS

This work was supported by the CONACYT, Mexico, under Grant No.: A1-S-47854 and by the Secretaría de Investigación y Posgrado, Instituto Politécnico Nacional under Grants No.: SIP 20200859, SIP 20200797, and SIP 20201948.

REFERENCES

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

Ayyaz Yaqoob Abid Rafiq Bashir Farhan, Ashraf Noman and Raza Ul Mustafa. 2019. Human aggressiveness and reactions towards uncertain decisions. *International Journal of Advanced and Applied Sciences* 6, 7 (2019), 112–116.

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *fifth international AAAI conference on weblogs and social media*.

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

Hugo L Hammer, Michael A Riegler, Lilja Øvrelid, and Erik Velldal. 2019. THREAT: A Large Annotated Corpus for Detection of Violent Threats. In *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*. IEEE, 1–5.

Sepp Hochreiter. 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6, 02 (1998), 107–116.

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

Linda Camp Keith. 1999. The United Nations International Covenant on Civil and Political Rights: Does it make a difference in human rights behavior? *Journal of Peace Research* 36, 1 (1999), 95–118.

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1746–1751. https://doi.org/10.3115/v1/D14-1181

Varada Kolhatkar and Maite Taboada. 2017. Constructive language in news comments. In *Proceedings of the First Workshop on Abusive Language Online*. 11–17.

Xuan-Hien Le, Hung Viet Ho, Giha Lee, and Sungho Jung. 2019. Application of Long Short-Term Memory (LSTM) Neural Network for Flood Forecasting. *Water* 11, 7 (2019), 1387.

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

Nelleke Oostdijk and Hans van Halteren. 2013a. N-gram-based recognition of threatening tweets. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 183–196.

Nelleke Oostdijk and Hans van Halteren. 2013b. 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*. 1034–1041.

Juan Ramos. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, Vol. 242. Piscataway, NJ, 133–142.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*. Association for Computational Linguistics, 19–26.

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

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 shared task on the identification of offensive language. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*. Vienna, Austria, 1–10.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*. 1391–1399.