A Review on Threat Detection Approaches in Social Networks

Ghadeer Al-Turaif[1] and Fethi Fkih[1,2],

[1]Department of Computer Science, College of Computer, Qassim University, Buraydah, Saudi Arabia
[2]MARS Research Laboratory, University of Sousse, Sousse, Tunisia

Abstract
The massive amount of data residing in social networks, becomes a fertile source of much relevant knowledge. Violent and criminal language is a very important knowledge that can be extracted from tweets. It is very interesting to predict suspicious threat language that can threaten the privacy or the integrity of a person or a community. Threats posted via social networks has possible to cause suffering and harm on an individual and society. Systematic studies of threat from a computer science perspective, is still recent. This paper will present a related works on automatic threat detection, including algorithms, methods, and text analysis features used. Additionally, we introduce a threat dataset consisting of 2440 tweet messages in English. Each tweet is manually annotated as either being a Threat or Non-Threat. The threat dataset is useful for training algorithms such as machine learning and deep learning, and for studying the nature of using threat words in society. This paper also discusses challenges of automatic threat detection. The development of shared resources, such as annotated datasets, algorithms and open-source code and platforms is a very important step to advance the automatic threat detection.

Keywords:
Threat; Social Networks; Annotated dataset; Text analysis.

1. Introduction

Nowadays, social media is playing an essential role in our lives as a platform for knowledge creation and sharing. Twitter became one of the popular social networking sites as it is an affluent area of much useful information [1], such as sentiment, opinion, trends, etc. Unfortunately, some of the discussions may be contaminated by offensive behavior like making violent threats [2] a development which is a concern [3] and threatens the privacy or the integrity of a person and community.

Moreover, there is a dark side to the internet as to the perceived anonymity of people being harassed, intimidated, and threatened [4]. Intimidator exploits the use of social media communication for threatening people. This threat can be driven by rage, revenge, or wanting to control others and feel stronger. So, social media providers thus struggle to provide better services to their users [3]. This leads to analyzing Twitter data for awareness of threats, but it is challenging to make it manually [5]. Due to the massive amount of data residing on Twitter, it will be impossible to detect threats manually. Hence, it becomes necessary to design automatic and efficient threat detection techniques, to help protect society and make necessary procedures against these people.

According to the international law in (Article 20)[1], it prevented any posting threat on the internet. Furthermore, most countries have strict national rules to reduce cybercrime, including threats. Additionally, some people have been arrested and prosecuted for posting threats or harming [6] under cybercrime laws.

As a contribution to solve this challenge, we present a new dataset of tweet messages in English language, where each tweet (manually labeled) is annotated as either being Threat or Non-Threat. As far as we know, the majority of the existing dataset of threats are very small and contain a small part of threats [3]. This is considered a weakness in the knowledge modeling domain which can negatively impact many fields of research, for instance, information retrieval, sentiment analysis, knowledge extraction, etc. In this paper, we introduce our dataset of threats extracted and collected from Twitter. Furthermore, we also contribute to give a solution for this problem by providing an overview of research conducted in this area and its challenges. Also, we present the problem, its definition, and identify approaches and resources that used.

After this introduction, in section 2 presents a theoretical background. Next, the previous related works of threat detection in social networks will be reviewed. We also show a summary including qualitative data (e.g., text analysis techniques in previous works) in section 3. Section 4, we explain the way of collecting data with description of dataset and analysis it. Then, challenges and discussion of automatic threat detection will be presented in section 5. Finally, the conclusion is presented in section 6.

2. Background

2.1 What is threat?

[1] United Nations Human Rights https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx

Manuscript received October 5, 2021
Manuscript revised October 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.10.49

Threats are a textual or verbal practice that implicates any violence against individuals regarding their ethnicity, gender identity, nationality, religion, personal conflicts, etc. In the same context, it is language meant to make the target (i.e. person under threat) or a broader group feel scared or unsafe [7]. This violence may be against a person, either specific or anonymous, or a group of people, such as feminists, black and white people, or also, it can be an explicit threat or an implicit threat. The explicit threat is threatening using threat words such as "I will hit my sis". The implicit threat is threatening, using a set of words that mean threaten. For example, "Blood washes with blood", which means "I will kill you".

2.2 Machine Learning and Deep Learning

Machine learning and deep learning are applications of artificial intelligence (AI). Deep learning is considered as a sub-field of machine learning as shown in Fig. 1 [8]. The machine learning domain come out from traditional statistics and AI communities [9]. Machine learning can be defined as a computational approach using available data and experiences to improve performance or to make accurate prediction [10]. Machine learning is the science of algorithms that relies on data and that make it the most excited field in all computer sciences. Machine learning algorithms are self-learning algorithms; therefore, they can convert solid data into knowledge. In the same context, machine learning refers to computer algorithms that ability to learn itself from available data rather than being totally programmed to solve a problem [11]. Thus, like human brain learning, the computer comes to be able to learning and enhancing its performance from acquired knowledge [11].

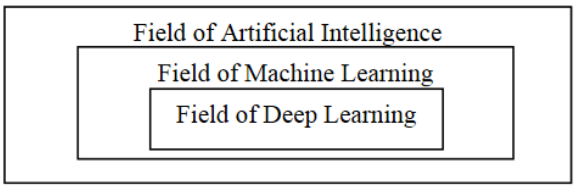

Fig. 1 Scope of machine learning and deep learning in AI [8].

There are several machine learning and deep learning algorithms are available, we will list some algorithms used in the context of threat detection.

- ***Support Vector Machine (SVM):*** the most common classifier used in text classification is the support vector machine (SVM) [12]. Also, SVM is the best-supervised machine learning classifier and gives the most accurate results in text rating issues [13, 14]. The goal of SVM is to find a function in a multidimensional space capable of separating training data from known class labels [15].
- ***Random Forest (RF):*** is a model that is utilized widely in classification [16]. The random forest model has a collection of tree predictors (decision trees), which train many trees in parallel and take the final decision of the RF model using the majority decision of trees [16].
- ***Multilayer perceptron (MLP):*** is a complement of feed forward neural network [17]. It is used to solve the classification tasks and often used in supervised learning tasks [18]. It comprises of three types of layers: input layer, output layer and hidden layer [17].
- ***Logistic Regression:*** is a statistical classifier derived from linear regression [19]. It is a model that creates a linear relationship between one or many independent variables and one dependent variable. Logistic regression is used to predict the probability of an outcome of the sigmoid function that only has two values, either 0 or 1.
- ***Convolutional Neural Network (CNN):*** is a specific kind of feed-forward neural network algorithms [20]. It is a deep neural network, which has a multilayer neural network including convolution and totally connected layers [21].
- ***Long short-term memory (LSTM):*** is a particular type of recurrent neural network with a strong capability to learn and forecast sequential data [22]. It is essentially a type of recurrent neural network (RNN) architecture [23], RNN has limited in preserving long-term memory [22]. Thus, the LSTM has ability to overcome this restriction by adding memory structure to preserve its state in the passage of time [22].

3. Related Work

In this section, we throw a spotlight on most important works that used Text analysis for detecting threats in textual data. There are two types of analyzing texts: qualitative and quantitative. Linguistics is known to be a qualitative approach and statistics is a quantitative approach. We divided it into three subsections: statistical, linguistics and hybrid approaches. Hybrid approach combines linguistic and statistical techniques.

The dataset of YouTube comments developed by [3] is considered as an important corpus in English language, since it was used in many research projects such as: [24, 25, 6]. This dataset (comments) was collected from nineteen YouTube videos about religious and political topics in 2013. Sentences were manually annotated that

included a threat of violence (or sympathy with violence) or not.

3.1 Statistical Approach

Authors in [24] performed a comparative study between several machine-learned models in order to automatically detect violent threats from a corpus of YouTube comments. For the experimentation phase, they used the dataset mentioned in [3] and the corpus was manually divided into sentences. To achieve this goal, the authors applied four different kinds of features to the corpus of data: linguistic, lexical, morphosyntactic, and semantic. The proposed model used three classifiers: Maximum Entropy (MaxEnt), Support Vector Machines (SVM), and Random Forests (RF).

As an extension of [24] authors in [25] used deep learning-based techniques, particularly convolutional neural networks, and GloVe (is an unsupervised learning algorithm for obtaining representations of vectors for words) for the word embedding task.

The work described in [6] used text mining and machine learning techniques to automatically classify YouTube's comments in order to detect threats or sympathies with violence. They used the available dataset provided by [3] and manually denoted as violent threats or not. Then, they made some preprocessing on the textual content. They used different feature matrices for comparison purposes, such as the document term matrix, bigrams of important words, and bigrams of important words with weight function. The classification is done using logistic LASSO regression if the sentence is violent or not.

In [18] introduced a new dataset for detect threatening language in Urdu tweets with implement many experiments in machine learning and deep learning classifiers. The proposed dataset is publicly available in GitHub repository[1] and manually annotated as threatening and non-threatening. Also, threatening tweets were classified into two categories: threatening for an individual or threatening for a group. Machine learning classifiers are Logistic Regression (LR), Multilayer Perceptron (MLP), Ada-Boost, Random Forest (RF), and (SVM), also, deep learning classifiers are 1-Dimensional Convolutional Neural Network (1D-CNN) and (LSTM). Preprocessing and tokenization of tweets. Features used fast-Text pre-trained word embedding and extracted n-grams for word and character using TF-IDF weighting.

3.2 Linguistics Approach

In [26] proposed a linguistic method to classify Arabic tweets to three classes of harassment which are: Terrorism, Violence and Threat. They divided their methods to three general steps. First step, they used a collection of linguistic resources for Arabic language to annotate their corpus. It consists of 1,998 tweets for training and 294 for testing. Linguistic resources are the Electronic Dictionary for Arabic "El-DicAr", the grammars of Arabic Named entities recognition and the grammar of segmentation elaborated by [27]. Second step involved of determination of linguistic patterns for each class, terrorism, violence and threat. They built a list of trigger words and extracted synonyms for each trigger using the Arabic Wordnet. These patterns written into local grammar. The third step is to convert these patterns into a set of transducers using the linguistic platform NooJ.

3.3 Hybrid Approach

Authors in [28] investigate approaches that use deep-learning algorithms to detect threats of violence and classify them based on whether they are directed at individuals or groups in YouTube comments. They used the available dataset [3]. The dataset is labelled into two categories by using majority voting. Deep-learning classification algorithms used are 1D Convolutional Neural Network (1D-CNN), Long short-term memory (LSTM), and Bidirectional Long short-term memory (BiLSTM) with different kinds of features, namely, a bag of words (BOW) with term frequency-inverse document frequency (TF-IDF), GloVe4 and fastText5.

The study presented in [29] presented a system for detecting terrorist threats on Twitter using a supervised machine learning algorithm. They used a dataset, available on kaggle.com, concerning Islamic State (ISIS) supporters that occurred in November 2015 in Paris. Tweets are classified manually into either a threatening class or a non-threatening class. They applied a sequence of tasks to the data, such as tokenization, lemmatization, stop word elimination, and finally, the processed data will be converted into numeric vectors. The proposed model uses a Support Vector Machine (SVM) algorithm for classifying a given text as containing a threat or not.

Authors in [4] discussed two approaches for the detection of threats in Dutch tweets. They downloaded tweets from the www.doodsbedreiging.nl website. They prepared the data by cleaning it up, like removing hashtags, retweet symbols, and URLs. They converted usernames to lowercase. The first approach is an n-gram pattern that is manually constructed, namely, unigrams, bigrams, trigrams, and skip grams (using two n-grams together, such as bigrams and trigrams). There is a lot of spelling variation in tweets, so it is possible to include all possible spelling variants of each word. The second approach is to use machine learning to determine tweets that have a threat or not, using token n-grams.

[1] https://github.com/MaazAmjad/Threatening_Dataset

Another system, proposed by [30] to detect threatened Dutch tweets based on trigger keywords and contextual cues. The pipeline of threat detection composes a preprocessor and two classifiers. After preprocessing tweets, in first classifier constructing a threat triggers list in a semi-automatic way from training dataset. In this stage, all tweets are lemmatized and computing Matthews correlation coefficient (MCC). Then, the outcome list of lemmas was manually cleaned, ordered by their correlation of threatening tweets class. Second classifier was applied and compared in two different approaches; one based on contextual cues for the triggers and another one based on patterns for the triggers. The Context approach works on mining of positive and negative cues for the triggers and Pattern approach uses alignment technique to mine threat patterns from training dataset, by using Needleman-Wunsch (N-W) algorithm. Three datasets used one for training and two for testing and evaluation. Training dataset is available in website (doodsbedreiging.nl).

The researchers in [31] presented an automatic system to detect threat and abusive languages in Bengali language from Facebook. Classification framework consist of tokenizer, preprocessing, classifier. Unicode Bengali words and emoticons are considered as valid input. Besides, consecutive exclamation and question marks are taken consideration. Classifiers used are Multinomial Naïve Bayes (MNB), Support Vector Machine (SVM) and Convolutional Neural Network (CNN) with Long Short Term Memory (LSTM). TF-IDF for unigram, bi-gram and tri-gram implemented for MNB and SVM classifiers. Then, CNN-LSTM is designed with four layers.

The researchers in [32] constructed a model to classify Instagram content (pictures and Arabic comments) to detect threat. The model was constructed by deep learning algorithm, specifically, CNN algorithm. Instagram dataset was collected and classified manually from various accounts on various subjects. Pictures and Arabic comments dataset were classified as threat and non-threat. For Instagram comments, researchers applied Arabic text preprocessing. TensorFlow framework was used in this model. TensorFlow is an open-source platform that works in heterogeneous environments developed by Google [33].

Finally, the work of [34] explored ways to analyze and extract threats from non-threats from Twitter. The Tweets were obtained from the Social Sentinel Company and classified as threats or non-threats by a computer before being reviewed by humans. They used a keyword in context (KWIC) tool to analyze and filter out data. The process of categorization is divided into three steps. These steps contain POS tagging, syntactic processing, and domain analysis. The result from these steps is eighteen categories, depending on use, such as golf, football, adjectives, etc. The initial analysis was to obtain the different uses in the dataset and capture small categories that can be combined. These categories were joined into ten thematic categories, such as sport. Then, they calculated mutual information scores, term frequency and statistical tests based on collocation to identify the most significant words and significant relationship between two collocates.

3.4 Summary and Analysis

The follow table introduces a summary of all papers previously are discussed. Table 1 can give as a quick reference for works that performed in the automated threat detection in social media.

4. Building threat dataset

In this section, we describe the dataset and process that we followed to build the first English dataset (i.e., from Twitter) for detecting threat language. We also provide the statistics of the resulting threat dataset. Finally, we present analysis to understand of threatening language that used in Twitter and the challenges.

4.1 Data Collection

Due to the lack of dataset dedicating to study threats in Twitter in English language, we got going to build one by ourselves. In this section, we will describe the methodology used for elaborating threat datasets. Twitter Application Program Interface (Twitter API) allows applications to communicate with each other to demand and deliver information. Twitter makes it easier for academics, researchers, and business developers to build their datasets using the Twitter API. Standard APIs provided by Twitter gain access to the data only for the past seven days, a small amount of the total volume of tweet messages. Regrettably, Twitter makes limits on the developer's account to provide reliability. Rate limits for calls to the API are different based on the authorization method that is used. For example, OAuth calls are allowed to 450 requests per 15 minutes[2].

The Twitter data used in this research was obtained in two ways: from the Twitter Stream API based on a set of threat keywords or hashtags, and by downloading available datasets. In this phase, we used the R programming language using RStudio software and the "twitteR" package that delivers access to the Twitter API to obtain tweet messages. We collected tweet messages using the search of threat-related keywords and hashtags, such as kill, hostage, attack, etc. We gathered approximately 355,000

[2] Document of rate limits, https://developer.twitter.com/en/docs/rate-limits.

Table 1: Summary of the threat detection works.

Paper	Language	Source of data	Text analytic Techniques		Classification algorithms
			Statistical	Linguistics	
[24]	English	YouTube comments	▪ Brown cluster.	▪ Lemmatization, ▪U niversal POS, ▪ Penn Treebank POS-tags, ▪ Dependency Relation, ▪ WordNet.	MaxEnt, SVM RF.
[25]	English		▪ Document term matrix, ▪ Bigrams, ▪ Bigrams with distance function.	-	Logistic LASSO
[28]	English		▪ TF-IDF weight, ▪ Features Extraction by GloVe and fastText, ▪ Bag of words features.	-	1D-CNN, LSTM, BiLSTM.
[4]	English	Twitter	▪ Numeric vectors.	▪ Lemmatization, ▪ Tokenization.	Simple Vector Machine
[29]	English	Twitter	▪ KWIC, ▪ Frequency words, ▪ Mutual information.	▪ Tokenization, ▪ POS tagging, ▪ Syntactic structure such as adjective, ▪ Transform variations of the data to one form.	-
[4]	Dutch	Twitter	▪ N-grams.	▪ Tokenization, ▪ Spelling Variation.	Machine learning system
[31]	Bengali	Facebook	▪ TF-IDF, ▪ N-gram.	▪ Tokenization ▪ Replace consecutive exclamation and question marks, ▪ Bengali stemmer.	MNB, SVM, CNN-LSTM.
[18]	Urdu	Twitter	▪ N-gram, ▪ FastText.	-	LR, MLP, RF, SVM, 1D-CNN, LSTM.
[30]	Dutch	Twitter	▪ MCC, ▪ The threat score, ▪ (N-W).	▪ Tokenization, ▪ Lemmatization	-
[32]	Arabic	Instagram	▪ Word2vec.	▪ Tokenization.	CNN
[26]	Arabic	Twitter	-	▪ Electronic Dictionary for Arabic “El-DicAr”, ▪ Arabic Named entities recognition, ▪ Segmentation elaborated by [27], ▪ Arabic Wordnet, ▪ NooJ platform.	-

unique tweets over two months and seven days in September, October, and November 2020, respectively. We have obtained variant information from tweets, relevant and irrelevant, such as news, public opinions, movies, and sports. There was particular interest in the content of tweet messages containing threats. Additionally, we mention four datasets available online which have been downloaded. The first dataset by [7] is about violent online harassment on Twitter. The dataset includes 35,000 tweets that they labeled as "harassing" or "non-harassing." The second dataset by Fifth Tribe1 was the terrorist attack by Islamic State (ISIS), which occurred in Paris during November 2015 and comprises 17000 tweets. The third and fourth datasets about suicide are published on GitHub2 and [35]. We selected a small part of the above datasets that contained threats.

Furthermore, we collected tweet messages for the test dataset of our project. The test dataset was collected and ultimately ended up with approximately 8775 tweets over 12 days, between June 18 and June 30, 2021.

To build training and test datasets, we dropped duplicate tweet messages, also non-English, using the R Language. After that, we read tweet messages by a human annotator (authors) to extract threat tweet messages from all datasets. Tweet messages are labeled as Threat or Non-Threat. Table 2 shown examples of threat and non-threat tweet messages. Finally, we will describe it in detail in the next subsection.

Table 2: Sample tweet messages of threats and non-threats.

Tweet messages	Class
If i do n't die by suicide or old age, my allergies will kill me	Non-threat
I will kill @user them	Threat
@user plz someone kill this guy	Threat
That salad pasta was bomb	Non-threat
@user i would but then whos gonna kill humanity?	Non-threat
14 killed, 75 Wounded in Bomb Attacks in South Philippines via @user | #explosives #bomb #attacks #violence #terrorism #terrorist #IslamicState #highthreat #threat #publicsafety #response	Non-threat

4.2 Dataset Description

The training dataset consist of 2240 tweet messages, which contains 1003 threat tweets and remaining is 1237 non-threat tweet messages. The test dataset contains 200 tweets: 100 threat tweets and 100 non-threat tweets. Then, we are compiling training and test datasets to be one dataset, called Threat dataset. Table 3 describes the numbers of tweet messages for training and test dataset.

Table 3: The number of tweet messages in Twitter threat datasets

Dataset	**Threat**	**Non-Threat**	**Total**
Training	1003	1237	2240
Test	100	100	200
Threat	1103	1337	2440

Threat dataset was collected from various events, such as the presidential election in the US and racial and ethnic conflicts. The dataset contains several types of threats targets which can be the owner of the tweet himself, like suicide threats, a person like threats to kill or hurt someone, a place, etc. Also, the target can be broad, such as "blacks," or specific, such as a known individual. Table 4 presents four examples of tweets that contain different types of threat targets.

Table 4: Examples of threat tweets.

Id	Tweets	Target
1	@user plz someone kill this guy (URL)	This guy
2	@user 😡😡😡 i will kill someone	Someone
3	I have a specific plan to kill myself	Myself
4	I will kill all the blacks tonight, tomorrow and any other day if they go to (name) university	Blacks

4.3 Analysis of Threat Dataset

Beyond its usefulness in understand of the natural language, there is also important to know patterns of threat used. In fact, words of threat take diverse meanings either in a threat sentence or in a usual sentence that talked about sentiments, opinions, etc.

In what follows, we try to provide a linguistic specification of the semantic ambiguity that can be induced from many threat vocabularies. This linguistic specification shed light on the difficulty of identifying the true sense of a sentence containing threat words.

Examples 1 and 2, indicate that a word can take many different and contradictory meanings. For instance, 'kill' was used in example 1 to threaten someone and in example 2 it expressed a feeling.

Example 1: @user plz someone kill this guy (URL)

Example 2: RT @user: Announce the deal💁💁 The suspense is killing me 😭😭

1 The dataset is available via https://www.kaggle.com/fifthtribe/how-isis-uses-twitter.

2 The dataset available via https://github.com/reetika-goel/Predict-Suicidal-Ideation-Based-on-Tweets.

In the same context, 'kill' can be used in apart of actual meaning, as in the sentence "trusting too much kills you" which refers to an opinion.

Also, threat words can be used for talking about taste of food or when people talk about their skills of cooking. For instance, 'bomb' word in the following examples 3 and 4, is used to describe the sandwich as delicious and her/his skill for made salad.

Example 3: Breakfast sandwich bomb ♡ 😍
Example 4: Made some bomb Chicken Salad

The previous examples demonstrate that, rather than using these words for the purpose of threatening, they can be used to praise or vilify someone or something, such as food, life, shape, people, etc. Twitter users used 'bomb' and 'fire', for example, as synonyms of beautiful, better, delicious, etc., On the other hand, words like 'kill' and 'death', for example, can be used to attack or criticize someone or something instead of using it to express feeling bored, sick, pain, etc.

The linguistic phenomenon discussed above is commonly known in the NLP field as "semantic ambiguity," which occurs when a word has more than one sense. We have to mention that it is difficult to eliminate the semantic ambiguity phenomenon since it is a very variable field. Furthermore, language in social media is constantly changing [36]. In fact, the semantic ambiguity in social media is continuing to increase as time passes increases and the meaning of a given word can change with the change of people's lives, such as their contact with different cultures and ideas to carry new meanings compared to their past meanings [37, 38]. Perhaps one of the primary reasons behind this thing is the wide use of social media, which is a fast-medium for the spread of ideas and changing the meanings of words, as we mentioned in the examples above, especially in words of threat.

5. Discussion and Limitations

Through literature review, we founded that there are not a lot of papers published in automated threat detection in textual data from a computer science, exactly in AI perspective. Most of researchers tend to collect and hand-code new data, and often these datasets be still private and not available in public repository. This reduce of the progress of researchers because limited availability of data, which makes comparison of results from various studies difficult. On another hand, it is hard to judge the effectiveness and performance for different features and classifiers, because each researcher used different datasets. However, we found three datasets are available, in English, Urdu and Dutch. We have reviewed the various studies for threat detection using text analysis techniques and algorithms. And due to the lack of datasets, we reach that there is no specific method proving to reach of the best results among the several papers. Additionally, the most of datasets used are imbalanced that make effects on the classifiers' performance.

Finally, we provided an overview on how the automatic detection of threat in textual data has developed through the past years. We identified opportunities and challenges form previous studies in this area, namely the rarely find of the open-source code and platforms that automatically classify threat. Threat detection is not just spotting of threat keywords, it is a difficult task that have many challenges. Unfortunately, this is a deep field have impact for the society and have also many of research challenges.

6. Conclusion

Social media has become a very important avenue for users to track news and express their opinions on a variety of events. Twitter is used extensively for different purposes. Twitter users post a huge number of tweets every day on various subjects like the news, casual chatter, etc. Thus, there has been a wider spread of harmful behavior such as threats, harassment, cyberbullying, etc. These conversations are not devoid of threats. In this paper, we present threat dataset of 2440 tweet messages (i.e., in English language) labeled as Threat or Non-Threat. We spent months for collecting, developing, and refining on threat dataset created to capture threat content. We suppose that the dataset contributes to train algorithms such as machine learning and as a source of data to understand, analyze and detect the threat phenomenon. Additionally, we provide an overview of social media automatic threat detection methods. We divide research according to text analysis techniques to statistical, linguistics and hybrid approaches. The future work will include incorporating the best text features that is able to text analysis to detect and classify threatening language in twitter. Additionally, we welcome partners that can be able to contribute to expand the size of the dataset to detect threatening language in English.

We have to mention that the proposed corpus is available from the authors, but restrictions apply to the availability of these data, which should be used under license, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Twitter. Researchers could require the data via email to 391200349@qu.edu.sa or ghadeer.a.t@hotmail.com.

References

[1] D. Alorini and D. B. Rawat, “Automatic Spam Detection on Gulf Dialectical Arabic Tweets,” in *2019*

International Conference on Computing, Networking and Communications (ICNC), 2019, pp. 448–452.

[2] N. Chetty and S. Alathur, "Hate speech review in the context of online social networks," *Aggress. Violent Behav.*, vol. 40, pp. 108–118, May 2018.

[3] H. L. Hammer, M. A. Riegler, L. Ovrelid, and E. Velldal, "THREAT: A Large Annotated Corpus for Detection of Violent Threats," in *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*, 2019, pp. 1–5.

[4] N. Oostdijk and H. van Halteren, "N-Gram-Based Recognition of Threatening Tweets," vol. 7817, 2013, pp. 183–196.

[5] S. Mittal, P. K. Das, V. Mulwad, A. Joshi, and T. Finin, "CyberTwitter: Using Twitter to generate alerts for cybersecurity threats and vulnerabilities," in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2016, pp. 860–867.

[6] H. L. Hammer, "Automatic Detection of Hateful Comments in Online Discussion," in *Lecture Notes of the Institute for Computer Sciences*, vol. 188, 2017, pp. 164–173.

[7] J. Golbeck *et al.*, "A large human-labeled corpus for online harassment research," *WebSci 2017 - Proc. 2017 ACM Web Sci. Conf.*, pp. 229–233, 2017.

[8] P. P. Shinde and S. Shah, "A Review of Machine Learning and Deep Learning Applications," in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, 2018, pp. 1–6.

[9] T. W. Edgar and D. O. Manz, "Chapter 6 - Machine Learning," in *Research Methods for Cyber Security*, T. W. Edgar and D. O. Manz, Eds. Syngress, 2017, pp. 153–173.

[10] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2018.

[11] S. Cohen, "Chapter 2 - The basics of machine learning: strategies and techniques," in *Artificial Intelligence and Deep Learning in Pathology*, S. Cohen, Ed. Elsevier, 2021, pp. 13–40.

[12] J. Salminen, M. Hopf, S. A. Chowdhury, S. Jung, H. Almerekhi, and B. J. Jansen, "Developing an online hate classifier for multiple social media platforms," *Human-centric Comput. Inf. Sci.*, vol. 10, no. 1, pp. 1–34, 2020.

[13] A. Muneer and S. M. Fati, "A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter," *Futur. Internet*, vol. 12, no. 11, 2020.

[14] R. Singh and V. Goel, "Various machine learning algorithms for twitter sentiment analysis," in *Information and Communication Technology for Competitive Strategies*, Springer, 2019, pp. 763–772.

[15] J. Dukart, "Basic Concepts of Image Classification Algorithms Applied to Study Neurodegenerative Diseases," in *Brain Mapping*, A. W. Toga, Ed. Waltham: Academic Press, 2015, pp. 641–646.

[16] S. Misra and Y. Wu, "Chapter 10 - Machine learning assisted segmentation of scanning electron microscopy images of organic-rich shales with feature extraction and feature ranking," in *Machine Learning for Subsurface Characterization*, S. Misra, H. Li, and J. He, Eds. Gulf Professional Publishing, 2020, pp. 289–314.

[17] S. Abirami and P. Chitra, "Chapter Fourteen - Energy-efficient edge based real-time healthcare support system," in *The Digital Twin Paradigm for Smarter Systems and Environments: The Industry Use Cases*, vol. 117, no. 1, P. Raj and P. Evangeline, Eds. Elsevier, 2020, pp. 339–368.

[18] M. Amjad, N. Ashraf, A. Zhila, G. Sidorov, A. Zubiaga, and A. Gelbukh, "Threatening Language Detection and Target Identification in Urdu Tweets," *IEEE Access*, vol. 9, pp. 128302–128313, 2021.

[19] J. Hefner and E. DiGangi, "Chapter 5. Ancestry Estimation," in *Research Methods in Human Skeletal Biology*, 2013, pp. 117–149.

[20] S. Shajun Nisha and M. Nagoor Meeral, "9 - Applications of deep learning in biomedical engineering," in *Handbook of Deep Learning in Biomedical Engineering*, V. E. Balas, B. K. Mishra, and R. Kumar, Eds. Academic Press, 2021, pp. 245–270.

[21] Y. Song and W. Cai, "Chapter 4 - Visual feature representation in microscopy image classification," in *Computer Vision for Microscopy Image Analysis*, M. Chen, Ed. Academic Press, 2021, pp. 73–100.

[22] K. K. Al-jabery, T. Obafemi-Ajayi, G. R. Olbricht, and D. C. Wunsch II, "4 - Selected approaches to supervised learning," in *Computational Learning Approaches to Data Analytics in Biomedical Applications*, K. K. Al-jabery, T. Obafemi-Ajayi, G. R. Olbricht, and D. C. Wunsch II, Eds. Academic Press, 2020, pp. 101–123.

[23] S. Hiriyannaiah, A. M. D. Srinivas, G. K Shetty, S. G M, and K. Srinivasa, "A computationally intelligent agent for detecting fake news using generative adversarial networks," 2020, pp. 69–96.

[24] A. Wester, L. Øvrelid, E. Velldal, and H. L. Hammer, "Threat detection in online discussions," in *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2016, pp. 66–71.

[25] C. E. Stenberg, "Threat detection in online discussion using convolutional neural networks," 2017.

[26] W. Elahsoumi, I. Boujelben, and I. Keskes, "Rule Based Method for Terrorism, Violence and Threat Classification: Application to Arabic Tweets," 2020, pp. 209–219.

[27] I. Keskes, F. Benamara, and L. Belguith, "Discourse Segmentation of Arabic Texts Based on Cascade Grammars," 2012.

[28] N. Ashraf, R. Mustafa, G. Sidorov, and A. Gelbukh, "Individual vs. Group Violent Threats Classification in Online Discussions," in *Companion Proceedings of the Web Conference 2020*, 2020, pp. 629–633.

[29] K. Bedjou, F. Azouaou, and A. Aloui, "Detection of terrorist threats on Twitter using SVM," in *Proceedings of the 3rd International Conference on Future Networks and Distributed Systems*, 2019, pp. 1–5.

[30] M. Spitters, P. T. Eendebak, D. T. H. Worm, and H. Bouma, "Threat Detection in Tweets with Trigger Patterns and Contextual Cues," in *2014 IEEE Joint*

Intelligence and Security Informatics Conference, 2014, pp. 216–219.

[31] P. Chakraborty and M. H. Seddiqui, "Threat and Abusive Language Detection on Social Media in Bengali Language," in *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, 2019, pp. 1–6.

[32] S. AlAjlan and A. Saudagar, "Machine learning approach for threat detection on social media posts containing Arabic text," *Evol. Intell.*, vol. 14, 2021.

[33] M. Abadi *et al.*, "TensorFlow: A System for Large-Scale Machine Learning," in *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, 2016, pp. 265–283.

[34] A. Beach, "'It's So Bomb': Exploring Corpus-Based Threat Detection on Twitter with Discourse Analysis," 2019.

[35] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of suicide ideation in social media forums using deep learning," *Algorithms*, vol. 13, no. 1, pp. 1–19, 2020.

[36] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing,"Diffusion of Lexical Change in Social Media," *PLoS One*, vol. 9, no. 11, pp. 1–13, 2014.

[37] Fkih, F., Omri, M.N, "Hidden data states-based complex terminology extraction from textual web data model", *Appl Intell*, vol. 50, pp. 1813–1831, 2020. https://doi.org/10.1007/s10489-019-01568-4

[38] Fkih, F. and Omri, M.N, "Information Retrieval from Unstructured Web Text Document Based on Automatic Learning of the Threshold." *IJIRR*, vol.2, no.4, 2012, pp.12-30. http://doi.org/10.4018/ijirr.2012100102

Fethi Fkih received his Ph.D. in Computer Science from Faculty of Economics and Management of Sfax, Tunisia, in 2016. He is a member of MARS Research Laboratory at the University of Sousse, Tunisia. He is currently working as an assistant professor in the College of Computer, Qassim University, Saudi Arabia. His research interests focus on Artificial Intelligence, Text Mining, NLP, Recommender System, Web Mining, Sentiment Analysis, Information Retrieval, Document Indexing and Semantic Web.