I. J. Computer Network and Information Security, **2023, 1, 39-53**
Published Online on February 8, 2023 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijcnis.2023.01.04

Threat Modelling and Detection Using Semantic Network for Improving Social Media Safety

Fethi Fkih*
Department of Computer Science, College of Computer, Qassim University, Buraydah, Saudi Arabia
MARS Research Laboratory LR17ES05, University of Sousse, Tunisia
E-mail: f.fki@qu.edu.sa
ORCID iD: https://orcid.org/0000-0001-8937-9616
*Corresponding author

Ghadeer Al-Turaif
Department of Computer Science, College of Computer, Qassim University, Buraydah, Saudi Arabia
E-mail: 391200349@qu.edu.sa

Received: 20 September 2022; Revised: 25 October 2022; Accepted: 29 November 2022; Published: 08 February 2023

Abstract: Social media provides a free space to users to post their information, opinions, feelings, etc. Also, it allows users to easily and simultaneously communicate with each other. As a result, threat detection in social media is critical for ensuring the user's safety and preventing suspicious activities such as criminal behavior, hate speech, ethnic conflicts and terrorist plots. These suspicious activities have a negative impact on the community's life and cause tension and social unrest among individuals in both inside and outside of cyberspace. Furthermore, with the recent popularity of social networking sites, the number of discussions containing threats is increasing, causing fear in various parties, whether at the individual or state level. Moreover, these social networking service providers do not have complete control over the content that users post. In this paper, we propose to design a threat detection model on Twitter using a semantic network. To achieve this aim, we designed a threat semantic network, named, ThrNet that will be integrated in our proposed threat detection model called, DetThr. We compared the performance of our model (DetThr) with a set of well-known Machine Learning algorithms. Results show that the DetThr model achieves an accuracy of 76% better than Machine Learning algorithms. It works well with an error rate of forecasting threatening tweet messages as non-threatening (false negatives) is about 29%, while the error rate of forecasting non-threatening tweet messages as threatening (false positives) is about 19%.

Index Terms: Semantic Network, Threat Detection, Social Media Safety, Cybersecurity, Knowledge Modeling.

1. Introduction

The advent of online social media has been a part of our everyday lives, as the number of users in social media is growing significantly. Recent statistics reported by Statista display that Twitter has more than 353 million users as monthly active who post over 200 billion tweets per year [1]. Consequently, Twitter is one of the most popular social networking sites which it is a rich area for a lot of useful information [2]. Furthermore, social media sites are considered as a perfect place to build their virtual world, where communicate with each other and post their daily activities, opinions, interests using multimedia like text data, photos, videos, etc. Some of Twitter posts might be controversial that may be caused to incitement to self-harm or harm others [1,3]. Sometimes, user communities in Twitter and launched campaigns against certain groups of people like Muslims, feminists, and white genocide which could potentially degenerate into physical violence.

Social media provides a large amount of information that is opening up a new prospect for knowledge used to understand the community or a whole world, to solve problems, to reason logically, to respond to crises, to make decisions, etc. Recently, there has been a set of research for detecting and combating undesirable behavior in social media like threats, harassment, etc. There are several methods proposed to solve this problem, such as machine learning and deep learning methods. In this work, we build a threat-focused semantic network for the presentation of threats. We integrate the proposed semantic network into our model for threat detection on Twitter.

The semantic network is a type of knowledge representation formalism. Knowledge representation (KR) is considered a method for knowledge representation that was created to allow intelligent models to exploit this knowledge

for many applications: learning, classification, prediction, etc. Additionally, it is an important field of Artificial Intelligence (AI), and the knowledge can be used to conclude a set of information about the problem. The semantic network was proposed for representing knowledge as a graph, provided by Quillian in 1969 [4]. It is a diagram structure used to represent knowledge in patterns of direct or indirect interconnected nodes and links [4]. It has a long history and opened up a foundation for knowledge modelling and representation. The semantic network comprises a library of semantic entities (e.g., concepts or phrases) and their semantic relationships, which are usually linguistically or statistically meaningful [5]. The semantic network defines the relationships between concepts and enables us to extract meaning from text based on nearby concepts' co-occurrence [6]. Semantic network analysis augments traditional content analysis by allowing researchers to derive more decadent interpretations of thematic identified in texts [7].

Briefly, the proposed threat detection (DetThr) model explores tweet messages to forecast the threats by using a threat network (ThrNet). It is constructed by analyzing threat tweet messages to extract relationships between concepts. The DetThr model reads tweet messages, pre-processes the data, identifies threat targets, and classifies tweet messages as a Threat or Non-Threat using our threat detection algorithm. This model will allow exploring threats from Twitter and developing a method to eliminate or report them. After this introduction, the remainder of this chapter illustrates the thesis problem, aim objectives, and thesis contributions.

The main contributions of this work are as follows:

- Generating ThrNet, the proposed semantic network, from the first training dataset.
- Proposing DetThr model for threat prediction from tweet messages.
- Carrying out an experimental evaluation of DetThr model on test dataset, demonstrating the highest performance that achieved on the classification threatened tweet messages.

After this introduction, the remainder of this paper illustrates as in in section 2 provides a review of the existing works for threat detection, also semantic network generation methods. Next, the dataset used in this work illustrates in section 3. In section 4, we introduce our proposed semantic network ThrNet. Then, section 5 presents DetThr, the proposed model for threat detection. Results of the proposed ThrNet and DetThr model are discussed in section 6. Finally, section 7 provides the conclusion and the future work.

2. Related Works

In this section, we provide a brief review of some well-known researches in this field, which is focused on the classification of threats using classical Machine Learning or Deep Learning. Also, we provide an overview of the main techniques used for the generation of this type of knowledge representation.

2.1. Threat Detetction

There are few previous works devoted to detecting threats in text. Previous works involved threat detection in many languages on many social media using several NLP, machine learning, and deep learning techniques. Wester [8], Ashraf [9], and Hammer [10] used the same dataset mentioned in [11] from YouTube comments in the English Language. Given that it has been included into so many research initiatives, this dataset is regarded as an important corpus. This dataset (comments) was gathered from 19 YouTube videos in 2013 that dealt with religious and political subjects and were categorized as posing a threat of violence (or showing sympathy for violence). Additionally, Twitter contents are used in [12] to detect terrorist threats by machine learning and [13] investigate ways to analyze threat text for threat detection.

However, recently a few studies have investigated threat detection in other languages. Amjad [1] addresses threatening language detection in Urdu tweets using machine learning and deep learning classifiers, Chakraborty [14] presents an automatic system for detecting threat and abusive languages in the Bengali from Facebook comment and AlAjlan [15] proposed a model to classify pictures and Arabic comments from Instagram using deep learning. In the Dutch language also, Oostdijk [16] and Spitters [17] present methods for detecting threats from Twitter messages.

2.2. Semantic Network

In this subsection, we present some methods and models for semantic network generation and their practical applications. Semantic Networks can be applied in many domains as politics, academics, medicine, economics, and industry. For example, a study of Hurricane Harvey and how government and emergency management (EM) organizations use Twitter to create crisis response strategies [18], Twitter posts related to HPV vaccine [19], and measles outbreak [20]. The process of constructing the semantic networks may differ from one research to another based on their requirements or goals. Generally, the first step in this process is the data gathering, which is a critical and important task, since it has a strong influence on the rest of the processes. Then, the textual data undergo a preprocessing procedure which intends to clean the data from noise. For instance, removing stop words or punctuation, correct errors of spelling, etc.

A semantic network is a representation of the meaning between concepts or words by the relationships (networks) [21]. Indeed, it exists several methods for constructing the relationship between words, for instance, by analyzing a large text and counting the frequency of words or their joint frequency in a statistical window (co-occurrences) or using

organized databases such as thesaurus [21, 22]. The quantitative network metrics help to understand semantic networks such as density to measure how connected nodes are in the network and centrality measures to help assist in determining who is "important" or “central” to a network like betweenness centrality, closeness centrality, and eigenvector centrality [23]. There are multiple software or tools to visualize the semantic network, for example, UCINET, NodeXL, Gephi and Protégé.

3. Twitter Threat Dataset

In this work, we used the threat tweet messages dataset[1] described in [24]. This dataset has 2440 tweet messages, which is divided into 1103 threat and 1337 nonthreat tweet messages. The threat dataset is split into training and test dataset. We split the training dataset into the first and second training dataset based on our requirements. In fact, we need dataset that contains only threat tweet messages to build the threat network (ThrNet) which is called, in this paper, the first training dataset that consists of 1003 tweet messages. The second training dataset has 1003 threat tweet messages and 1237 non-threat tweet messages; it is used for training machine learning-based models, which were classified into threat and non-threat classes. Additionally, the test dataset contains 200 tweets: 100 threat tweets and 100 non-threat tweets. This dataset was used to test our model (DetThr) and machine learning models. Table 1 illustrates statistics of the first, second training, and test dataset. Some examples of threat and non-threat tweet messages from the threat dataset are shown in Table 2.

Table 1. Threat Dataset Description.

Dataset	Threat	Non-Threat	Total
First Training	1003	–	1003
Second Training	1003	1237	2240
Test	100	100	200
Threat	1103	1337	2440

4. Proposed Semantic Network for Threat Representation: ThrNet

In this section, we introduce ThrNet our proposed semantic network for presenting threats on Twitter. We propose in this work an automatic method for generating a threat network from unstructured data. Figure 1 presents processes performed to construct the ThrNet.

Table 2. Examples of Threat and Non-threat Tweet Messages.

ID	Tweets	Class
1	hi everyone, I found this guy on Omegle and he is threatening me to viral my photos	Threat
2	Typos are going to end my life	Non-Threat
3	This is not August 4, this is today. These are dark, dark days. The fire has not been put out yet. Photo by @AP #Lebanon #Beirut https://t.co/V4n7	Non-Threat
4	this guy is threatening to kill my family what do i do	Threat

4.1. Data Preprocessing

Tweet messages in their basic form contain a lot of media and noise, like emojis, videos, pictures, etc. Preprocessing of tweet messages is a fundamental task for text classification and threat detection. In this section, we discussed the process for preparing tweet messages for further phases.

A. Data Cleaning

We preprocess tweet messages to enhance the quality of the data that makes it suitable for analysis tasks. This way is to remove or modify data that is usually not useful, unnecessary or hinders data analysis. First, tweet messages were being tokenized i.e. it to separate text as a token. Second, we removed noise data from tweet messages like emojis, URLs, and emails. Third, all tweet messages are converted to lowercase letters to avoid variation and occurrence that is caused by differences of lowercase/uppercase letters in the same words. Furthermore, there is no distinction between words in uppercase and lowercase characters. Fourth, special characters such as #, ’, % and user names mentioned after @ were also removed from tweet messages. Finally, we replace some contractions of pronouns with full forms for the further phase. For example, we replace i’ll→i will and ’em→them.

[1] Please contact the authors if you would like to obtain access to the threat corpus

B. Lemmatization

In language, we write words derived from another word called inflection. It is a variation of word forms such as tenses, numbers, and cases. For example, the inflectional of the past tense of verbs in English can either be regular forms (e.g., kick-kicked, kill-killed) or irregular forms (e.g., see-saw, eat-ate) [25]. In particular, in languages with higher variation in surface [26], lemmatization plays an essential role in preprocessing tasks.

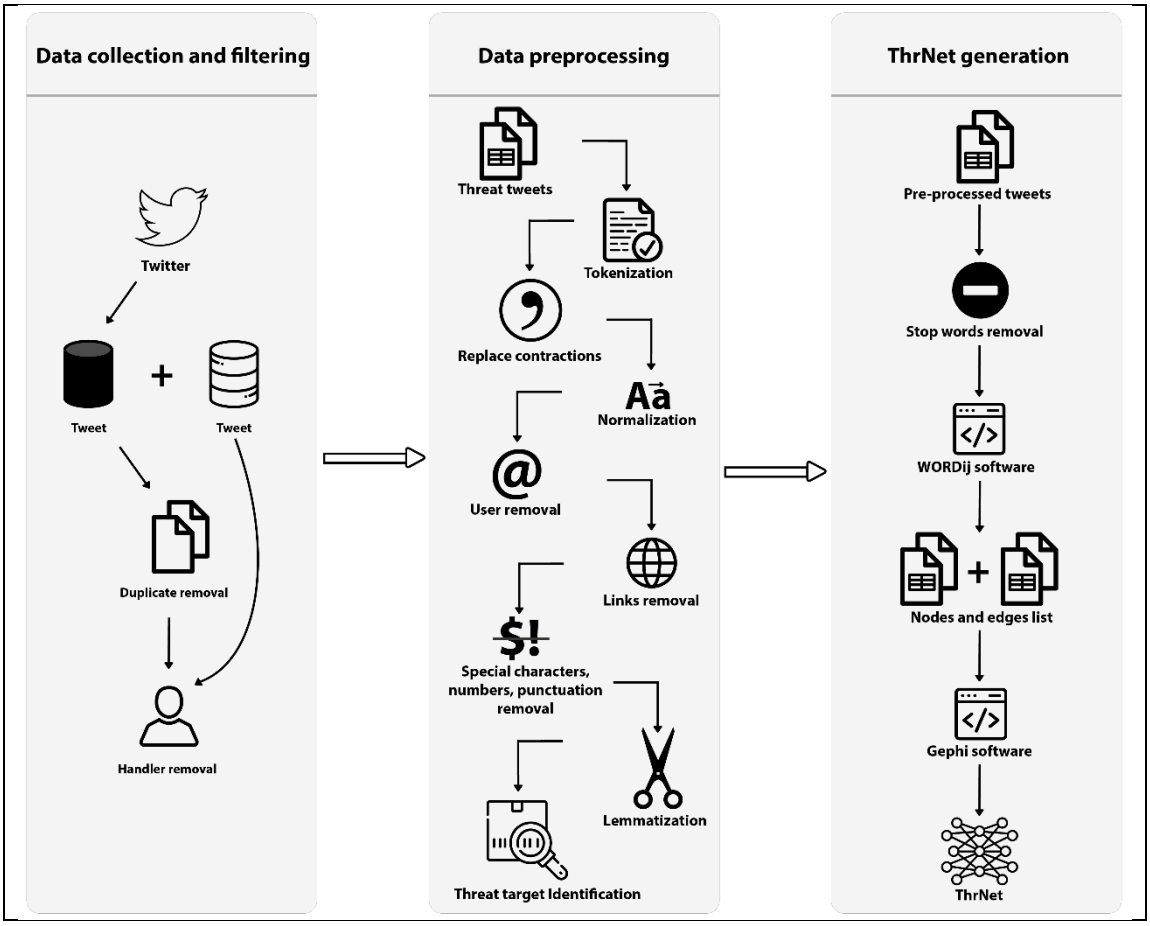

Fig.1. Steps of ThrNet Generation.

Lemmatization is a method of grouping variant words into one word by removing inflectional words [27]. This leads to making words from their base form or lemma based on a dictionary. Lemmatize words after fathom the Parts of Speech (POS) and putting a word in the given sentence [28]. It is one of NLP techniques, which is used to normalize text. The reason for using lemmatization is to reduce various forms of the same word in the ThrNet. We performed the lemmatization but it is not working well enough with the casual tweet messages language [29], for example, the word 'saw' didn't return to 'see'. Additionally, in a large dataset, lemmatizing words are damaged by avoiding information in conjugated forms [30]. We used NLTK[2] WordNet Lemmatizer for two training and test dataset. It is one of NLP techniques, which is used to normalize text.

4.2. Threat Target Identification

People posting their threats usually target an entity; therefore, identifying the entity is important. The identification of the target of the threat is a crucial task in the process of semantic network generation. For this reason, we use two techniques: Named Entity Recognition and phrase matching.

A. Named Entity Recognition

Named Entity Recognition (NER) is one of the NLP tasks to extract information from text such as Twitter, YouTube comments, emails, etc. A named entity is a word or phrase with similar attributes to identify one entity; for example, people, countries, and location names refer to LOC in NER [31]. NER is a process to identify named entities in text and classify them into predefined entity categories. In our case, we have an interest in recognizing threat targets in tweet messages because they help us detect threats. For the sake of extracting information from unstructured messages. NER

[2] https://www.nltk.org/. NLTK is a for NLP in library the Python language.

provides the basis for many natural language applications, such as machine translation, information retrieval, and answering questions [25]. NER is an online tool, such as in the R Language by the open NLP package and Python using the Spacy packages. We used a SpaCy [3] package to identify NER in our tweets and replace them with their labelled entities. NER aims to identify and classify mentions of rigid designators from the text that belongs to predefined semantic types like organization, time, person, etc [27,25]. It helps to recognize threat targets from tweet messages that may threaten a person or organization at a specified time.

B. Phrase Matching

Since the NER is not enough to determine all threat targets; we used rule-based matching, especially phrase matching5, to add a new target. We determine new threat targets that people used to threaten each other on Twitter. Because "target" is sometimes a pronoun, such as in the first example in table 2, "me" is the target. Also, in table 2, the fourth example, target, is "family", which denoted this group as "MEMBERS". We identify patterns that we want to match by adding these patterns to the Matcher tool and running it on tweet messages through the matcher to extract the matching phrases and replace them with their categories. The new target lists for each category are shown in tables 3 and 4. Also, Table 5 describes targets from NER and our threat targets.

Table 3. PRONOUNS Category.

My	She	He	You
They	We	I	My
Them	Us	Our	Me
Your	His	Her	Their
Anyone	Everyone	Everybody	Anybody
Someone	Yourself	Themselves	Herself
Himself	Yourselves	Ourselves	Myself

Table 4. MEMBERS Category.

Family	Women	Son	Great grandfather
Baby	Woman	Brother	Great grandmother
Kid	Young	Sister	Grandfather
Mom	Daughter	Sis	Grandmother
Dad	Lady	Bro	Grandson
Girl	Father	Husband	Great granddaughter
Guy	Mother	Wife	Uncle
Aunt	Cousin	Nephew	Niece
Man	Men	Male	Female

Table 5. Description of the Threat Target.

Target	Description
PERSON	Individuals, including fictional
ORG	Acronym of "organization", companies, universities, agencies, hospitals, institutions, etc.
GPE	Countries, cities, states.
FAC	Buildings, airports, roads, highways, bridges, etc.
NORP	Nationalities or religious or political parties.
PRODUCT	Items, goods, vehicles, foods, etc. (not services)
DATE	Bounded or relative dates.
TIME	Periods of time smaller than a day.
EVENT	Named storms, wars, sports events, political events etc.
LOC	Acronym of "location", mountain chains, bodies of water.
PRONOUNS	Personal, indefinite, reflexive, intensive pronouns.
MEMBERS	Names of family, gender.

[3] https://spacy.io/. SpaCy is an open-source library for NLP using Python language. [5]https://spacy.io/usage/rule-based-matching.

As seen in the original tweet, we can directly observe that the tweet contains noise that needs to be clean for analysis, then applies threat targets identification for replacing it with their entity. So that table 6 provides an example of the preprocessing procedure and the threat target identification.

Table 6. Example of Tweets Preprocessing.

The original tweet	"@user Hon. Member, Join the protest tomorrow and avoid the death of another Nigerian. You are a KEY stakeholder https://t.co/Fk3JLljcu#SARSMUSTEND
After cleaning	hon member join the protest tomorrow and avoid the death of another nigerian you are a key stakeholder
After Lemmatization	hon member join the protest tomorrow and avoid the death of another nigerian you be a key stakeholder
After Target identification	ORG member join the protest DATE and avoid the death of another NORP PRONOUNS be a key stakeholder
After stop words removing	ORG member join protest DATE avoid death another NORP PRONOUNS key stakeholder

4.3. ThrNet Generation

The semantic network is a map to represent concepts and relations between them that represent knowledge. This work employed the semantic network using the measurement method for text data, both statistically and graphically. The semantic network generates concepts from the tweet messages as nodes connected together by the frequencies with which each concept co-occurs with the other concepts [32]. It grasps the structure of co-occurring words or concepts, which allows it to understand the meanings that people create in their discussions [19]. The semantic network provides a way to analyze and extract information from texts [33]. We generate words, word pairs and their weights for the first training dataset using WORDij software[4] [34, 35]. It is a text analysis software that contends with concepts and words as nodes and word pairs as edges for network analysis [35].

Besides, WORDij computes the shortest paths between words using direct and indirect pair information [36]. Word pairs determine the link's strength as the number of times each word appears closely in text with another [35]. The first training dataset is analyzed in the WordLink module; it contains an option to remove a list of words that we used to drop stopwords, and this allows the removal of any set of words. Also, we have chosen the frequency of words and word pairs higher than three and excluded the lowest; we set the window size to five. Windowing stopped at the end of the tweet and appeared at the beginning of the next tweet. This construction of words and word-pairs frequencies was performed from the first training dataset. Then, the semantic network was constructed by connecting words based on co-occurrences. It is concluded that it plays an important role in generating relations in semantic networks. These processes produced the matrix in the pattern of an adjacency list of words and word pairs and their frequencies of occurrence. According to [37], larger window sizes reflect semantic concepts and other relations held over large ranges, but smaller window sizes define limited expressions and other relationships with short ranges. Window size 5 is large enough to show the limitations between verbs and arguments, but not so large to remove limitations that use strict adjacency. Finally, we are visualizing the ThrNet using Gephi software. It is an open-source program for the analysis and visualization of networks. We loaded the ThrNet data that consists of the list of words as nodes and the list of word pairs to link these words using Gephi software.

5. Proposed Model for Threats Detection: DetThr

In this section, we present the proposed threat detection model to predict whether tweet messages are threats or not. We are using the ThrNet in our proposed algorithm from tweet messages in the threat detection model. Figure 2 shows all the processes used for tweet messages using ThrNet's threat detection algorithm to predict Threat or Non-Threat tweet messages. First, we import the test dataset, which is classified as "Threat" or "Non-Threat". Second, we used the tweet-preprocessor[5] library in Python to preprocess tweets. This library is used to clean up the test dataset from URLs, mentions, reserved words (e.g., RT), emojis, etc. Preprocessor is a Python preprocessing library that is used for tweet messages. Third, we use the WordNet lemmatizer with NLTK, a process to convert words to their base form. Fourth, we identified threat targets in tweet messages using NER and phrase matching by SpaCy. Fifth, we remove stop-words with NLTK in Python. Finally, we are using ThrNet in the threat detection algorithm to predict whether a tweet message is a "Threat" or "Non-Threat".

As shown in algorithm 1 and figure 3, to decide if the target is threatening or not, we first look for the target in ThrNet and words connected to this target. In this case, at ThrNet, the target connects to four words; all of these words represent threats to the target. Moreover, on the tweet network, this is part of the sentence (tweet) that contains a target; to look for them in ThrNet and make a decision if they are a threat or not. First, it takes target with word3 and searches ThrNet to see if it can find it. Then it takes target with word5 or word3 with word4 and searches ThrNet to see if it is found; if it is, this tweet is a threat; otherwise, it is not a threat. To illustrate this with an actual example, from the test dataset (e.g. I am going punch (name) in the face), beginning look for the target in the tweet (i.e. I→pronouns) with word neighbour is as I→go→punch, then find it in ThrNet. Another example (e.g., Kill my mother too, please) is that "my mother" is the target (i.e. my mother→members) and is captured in the Threat as follows: kill→my mother→please. This

[4] https://www.wordij.net/.
[5] https://pypi.org/project/tweet-preprocessor/

shows how to find threats in targets and neighbours' words that link words together, according to the contexts of words.

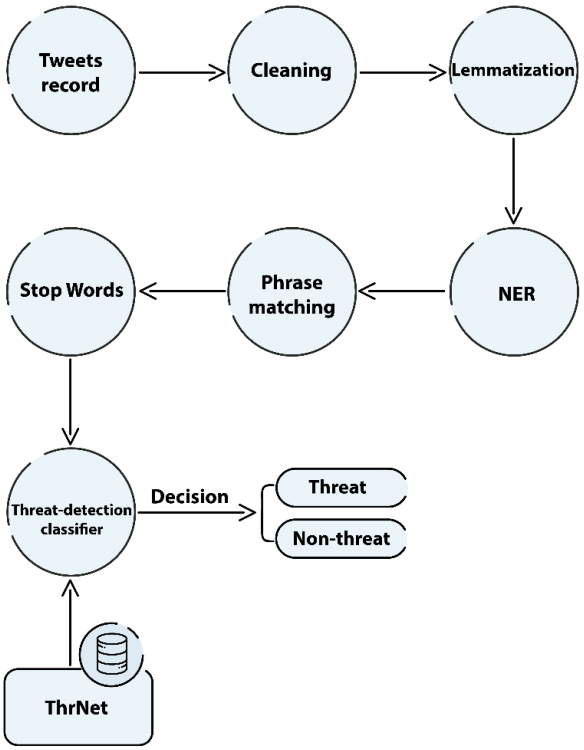

Fig.2. DetThr Model Preview Steps.

Algorithm 1 Threat detection

Require: A stream of tweets: Tweet: word1, word2, word4, word5, word6, ...
ThrNet: a threat network
Ensure: A prediction of classes
1: Target= all Target
2: **for** each Tweet **do**
3: **for** each word **do**
4: **if** word=Target **then**
5: X= word next Target
6: Y= word next X
7: Z= word previous Target
8: **if** Target and X = ThrNet OR Target and Y = ThrNet **then**
9: **if** X and Y = ThrNet OR Y and X = ThrNet OR Z and Target = ThrNet OR length(Tweet)= 2 **then**
10: Predict="Threat"
11: **else**
12: Xsynonyms= Synonym set of X from WORDNET [ ]
13: Ysynonyms= Synonym set of Y from WORDNET [ ]
14: Zsynonyms= Synonym set of Z from WORDNET [ ]
15: **for** Xsynonyms **do**
16: **for** Ysynonyms **do**
17: **if** Xsynonyms and Ysynonyms = ThrNet OR Ysynonyms and Xsynonyms = ThrNet **then**
18: Predict="Threat"
19: **end if**
20: **end for**
21: **end for**
22: **for** Zsynonyms **do**
23: **if** Zsynonyms and Target = ThrNet **then**
24: Predict="Threat"
25: **end if**
26: **end for**

```
                    end if
                else
                    Predict=”Non-Threat”
                end if
            else
                Predict=”Non-Threat”
            end if
        end for
end for
```

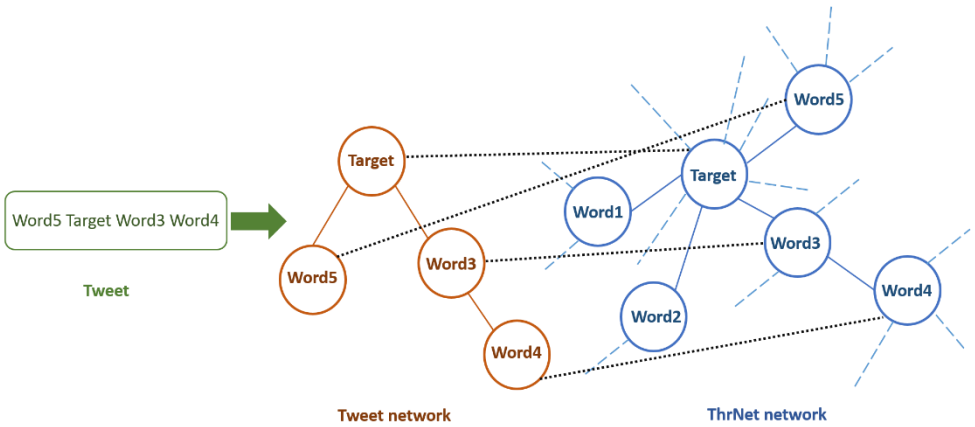

Fig.3. A Graphic Illustration of the Threat Detection Algorithm.

6. Results and Discussion

We discuss in this section the results provided by our proposed ThrNet and DetThr model. Additionally, we carried out a comparative study between DetThr and machine learning-based model.

6.1. Evaluation Metrics

For the evaluation phase we use the following measures:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

$$F1 - Score = 2 * \frac{Precision*Recall}{Precision+Recall} \tag{5}$$

Where:

- True Positive (TP): Number of threat tweets correctly predicted.
- True Negative (TN): Number of non-threat tweets correctly predicted.
- False Positive (FP): Number of non-threat tweets incorrectly predicted.
- False Negative (FN): Number of threat tweets incorrectly predicted.

6.2. Results

Our goal is to create a model that is capable of detecting threats from tweet messages using the semantic network. Thus in our experiment, we present our ThrNet and use it in DetThr model for threatening detection. Then, we will provide the results that were performed to DetThr model and machine learning-based model.

A. Semantic Network: ThrNet

The semantic network carries the structure of co-occurring words or concepts, introducing an understanding of the

meanings that people create and discuss with each other on Twitter [19]. In the same context, ThrNet helps to extract the meaning from the casual language that Twitter users post to each other. It captures different meanings from various conversations between users that were collected in the first training dataset. In fact, ThrNet is a graph of connected words that refer to the threat concept. A vertex or node can be a word belonging to the threat vocabulary or a target directly concerned by the Threat. Edges in the graph represent a statistical relationship between vertices: two connected words can occur in the same statistical context in the tweet message. Also, edges were constructed for words (concepts) that occurred within five words of each other inside each tweet. For example, the word "go" frequently occurs with some threat words like "kill", as shown in the example 'I'm going to kill myself', so it will be connected to this word in the graph. Also, we excluded all words that do not mean threats and have no effect on detecting threats. We generated a list of words and word pairs and their frequency of occurrence in the form of an adjacency list. The final co-occurrence list generated 521-word pairs ranging in frequencies between 3 and 196. Then, to visualize the ThrNet, we imported the words and word pairs list to the Gephi software. Also, the ThrNet was displayed using the ForceAtlas2 layout.

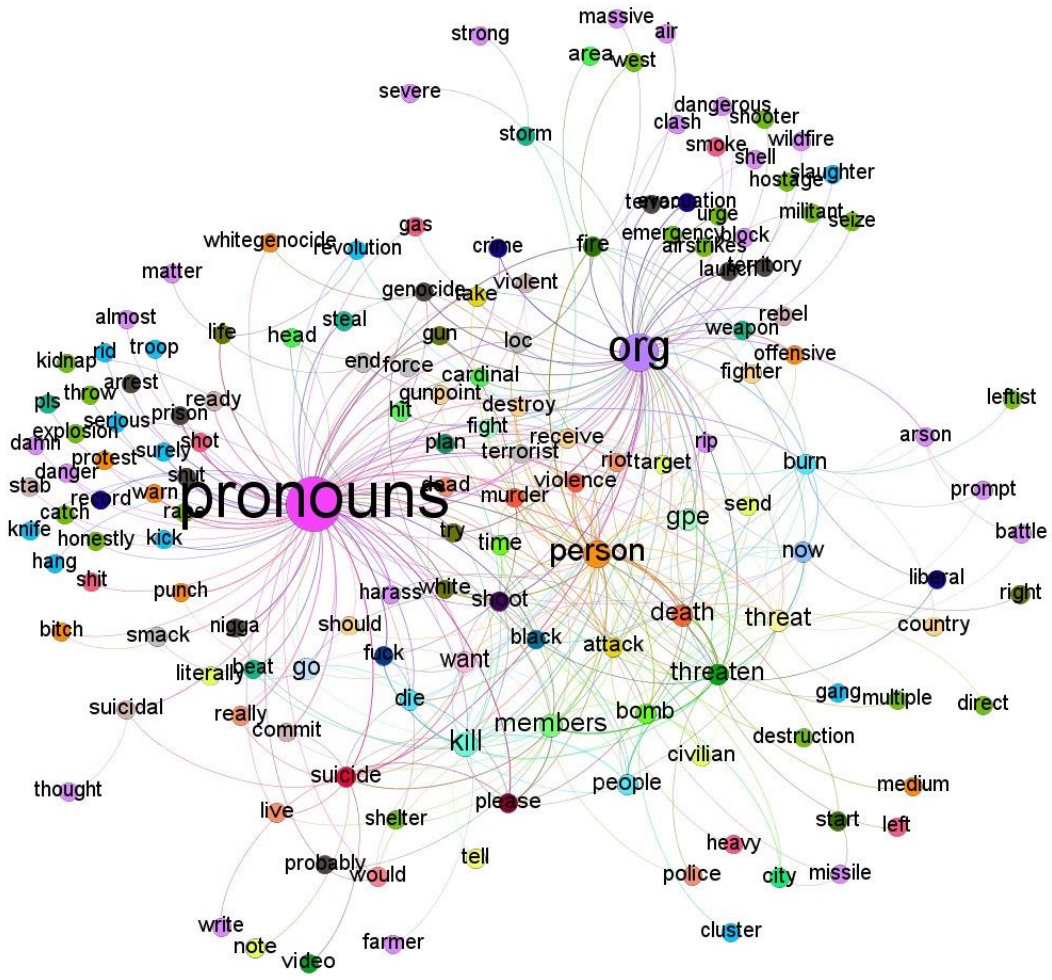

Fig.4. ThrNet, a Twitter Semantic Network for Threats.

The ThrNet contains a group of concepts and concept relations. Figure 4 visualizes the complete ThrNet graph generated from the first training corpus that contains the threat words and words they are connected. The ThrNet structure reveals a density network wherever most words are associated with targets. The density of the ThrNet is 0.032, meaning that 3.2% of the words in the tweet messages co-occurred with one another. Node size in ThrNet indicates the frequency of occurrence of concepts in the network. For example, pronoun targets appear more frequently than a person's targets, but both targets were linked directly or indirectly. The ThrNet reveals that some targets share the same threat words, such as 'murder', while some have their own threat words, such as 'kick'. Also, targets are connected to threat words or words related to Threat, whether direct or indirect, such as 'org' target link directly with the word "clash" and indirect with the word's "storm" and "strong".

That is, the closer the words were related, the shorter the distance between them. Furthermore, threat targets that have a frequency of less than five that do not appear in ThrNet, because they have don't interconnect with threat words over three times. For example, NORP target appears three times with three different threat words.

B. Testing DetThr Models

Machine learning techniques can be used to understand daily problems that people are facing and detect them automatically. The machine learning-based model is the most commonly used for text classification. The machine learning method is a form of text classification that uses a human-labelled training dataset to train a classifier [38]. The classification of text based on machine learning techniques can be referred to as a supervised learning problem. We chose the best performance of machine learning algorithms based on previous research on text classification such as threats, sentiment analysis, cyberbullying, and online harassment. We mention that Machine learning algorithms used in this work are Naive Bayes (NB), Support Vector Machine (SVM), Random forest (RF), Logistic Regression, Decision Tree (DT), XGBoost (XGB) and K- Nearest-Neighbor (KNN). We applied the same features to the machine learning-based model that are explained in sections 4.1 and 4.2. Machine Learning-based models are trained using the second training dataset, in which we classified tweet messages as threats and non-threats. We performed the test of the DetThr model and machine learning models using the same test dataset. The experimental study shows that the DetThr model outperforms the machine learning-based models. This result can be explained by integrating the ThrNet into the DetThr model, which has a positive effect on the model performance. We performed the test of the DetThr model and machine learning models using the same test dataset. DetThr model achieved a good accuracy of 76% and F1 score of 75%. The results of the comparative study between DetThr model and the Machine Learning models for threat detection from Twitter in terms of different evaluation metrics are shown in Table 7 and 7. We mention that NLP tools have a performance gap that affects the performance of the DetThr model and machine learning algorithms. Figure 5 illustrates the result of the DetThr model for two classes Threat and Non-threat. We generate confusion matrix from DetThr model using the test dataset. Figure 6 shows the confusion matrix for DetThr model.

Table 7. Result of DetThr Model and Machine Learning Algorithms for Threat Detection.

Model	Recall	Precision	F1- score	Accuracy
NB	55%	58%	52%	56%
SVM	61%	62%	61%	61%
RF	63%	63%	63%	63%
Logistic Regression	60%	60%	60%	60%
DT	53%	54%	53%	54%
XGB	55%	57%	54%	56%
KNN	55%	59%	48%	55%
DetThr Model	**71%**	**79%**	**75%**	**76%**

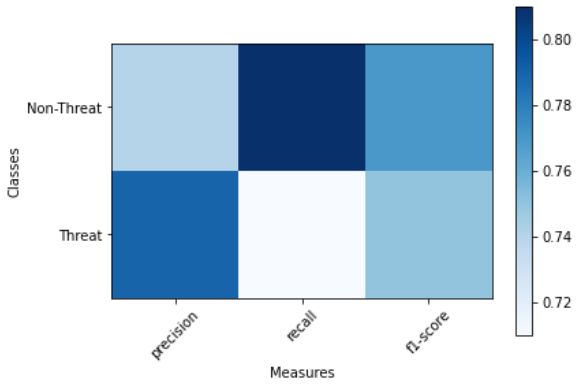

Fig.5. Result of DetThr Model.

C. Discussion

The evaluation results, illustrated in section 6.2, show that our proposed model surpasses the machine learning-based models. This result is proof of the large contribution of semantic resources to the performance of any application in the text analysis field. In fact, as a powerful tool for knowledge presentation, the semantic network can provide the model with better perception and understanding of the information to extract or retrieve, thereby increasing its performance.

However, the confusion matrix in Figure 6 shows that the false negative ratio is significant (29%), which indicates that the proposed model is incapable of recognizing an important amount of relevant information. Also, the false positive ratio is about 19%, which indicates that our proposed model can be infiltrated by a considerable amount of irrelevant information.

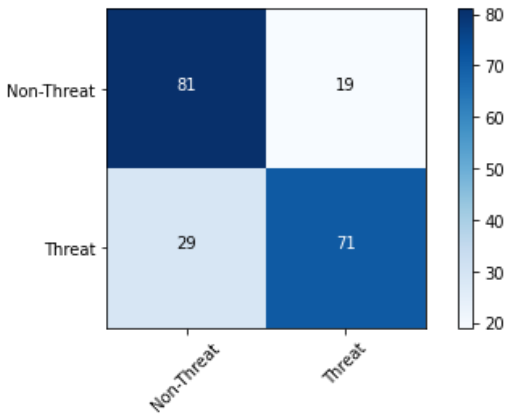

Fig.6. Confusion Matrix for the Threat Detection Model.

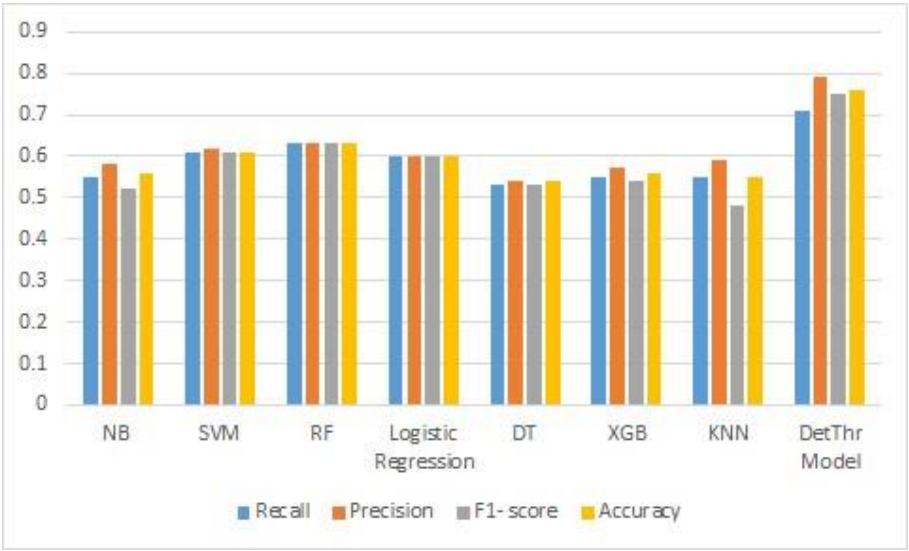

Fig.7. Performance Comparison between our Proposed Model and the Machine Learning Models in Terms of Recall, Precision, F1-score and Accuracy.

The obtained results are explainable by the complex structure of natural language, which can cause semantic ambiguity. In fact, a word related to "threat" vocabulary can have a variety of connotations, depending on whether it is used in a threat sentence or a language discussing feelings, opinions, etc. In the paragraphs that follow, we attempt to linguistically specify the semantic ambiguity that can be inferred from a variety of threat vocabulary. This linguistic distinction makes clear how challenging it can be to determine the genuine meaning of a sentence that contains threat words. The first and second examples show that a term can have a wide range of contradictory interpretations. For instance, the word "kill" was used to show emotion in example 1 and to threaten someone in example 2.

Example 1: *i Just Woke Up My Stomach Killing Me*
Example 2: *I will find you and I will kill your kid for that just wait for that*

In the same context, the word 'kill' can be used apart from its actual meaning, as in the sentence "@user 😂😂 bro boredom wan kill me 😂", which refers to talks about being bored. Likewise, threat words can be employed to describe the taste of food or to describe one's cooking abilities. In the following examples 3 and 4, the word "bomb" is used to describe the coffee latte as wonderful and his/her skills to make dinner.
Example 3: *This latte is really bomb today*
Example 4: *This is my favorite time of year to stay home and cook bomb ass dinners* 😍🥰

Moreover, influential people are posting and promoting these words on social media, such as marketing for products or describing their shape or life, as in the following examples 5, 6 and 7:

Example 5: *Serena is on fire in this one*🔥 *Serena's IG (URL)*
Example 6: *RT @user: Let life be beautiful like summer flowers and death like autumn leaves*
Example 7: *Maths is gonna be the death of me* 😔

Additionally, there are some tweet messages contain threats in the textual content but the attached media (images, videos, comments, quote tweets, etc.) clarifies the true meaning of the tweet which changes the meaning from threat to non-threat. Figure 8 shows an example of tweet containing an ambiguous textual content which is disambiguate with a video giving the clear meaning of the tweet a challenge in a game. Figure 9 also an example of a tweet with its comments that shows the last comment is threat in textual data, but the tweet illustrates that they are playing the online game.

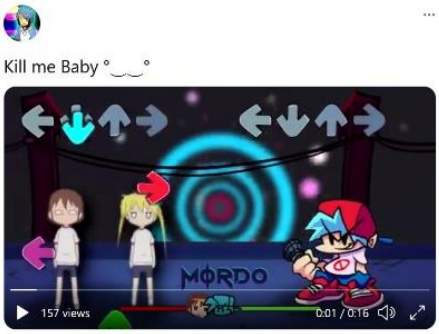

Fig.8. Example of an Ambiguous Textual Content in a Non-threat Tweet.

The previous examples show that, rather than being used to threaten someone or something, they can be utilized to praise or criticize someone or something, such as life, shape, food, people, and so on. Twitter users used words like ’bomb’ and ’fire’, as synonyms for delicious, beautiful, better, gorgeous, etc. Furthermore, words like ’death’ and ’kill’, can be utilized to assault or criticize someone or something rather than to describe feelings of boredom, sickness, suffering, etc. The language phenomena mentioned above is referred to as "semantic ambiguity" in the NLP discipline and happens when a term has more than one meaning [39,40]. This issue needs to be deeply handled in the future to increase the performance of our model, DetThr. In fact, semantic ambiguity can be resolved using many techniques, such as Word Sense Disambiguation (WSD) algorithms and linguistic resources (ontology, thesaurus, dictionaries, etc.) [41,42].

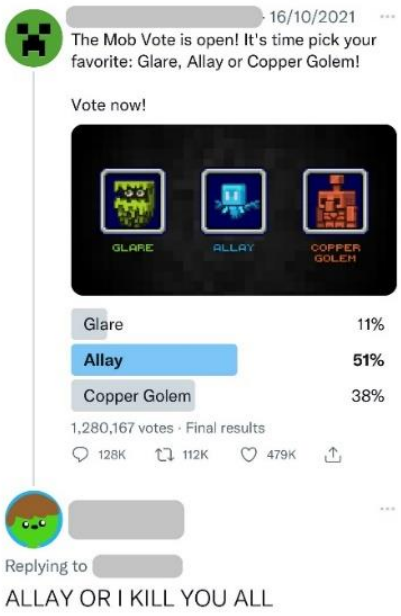

Fig.9. An Example of an Ambiguous Textual Content in a Non-Threat Tweet.

We should point out that it is hard to eliminate the semantic ambiguity phenomenon because it is a variable and developing field. Since the language in social media is ever-changing and evolving [43,44]. In fact, semantic ambiguity in social media is increasing as time passes, and the meaning of a given word might vary as people’s lives change, like their contact with various cultures and ideas, resulting in new meanings compared to their previous meanings. Maybe one of the first reasons behind this issue is that the widespread use of social media, which, as we saw in the examples above, is a fast-medium for the propagation of ideas and changing the meanings of words, particularly in words of threat. In the same context, rapid changes that occur in the present time, i.e. what’s applied today, is maybe invalid tomorrow. We mean that some word today has multiple meanings, could be those meanings changing in the future, whether some of them disappear or carry other meanings. Senses of words are produced with changing times or trending social media, some of these meanings disappear with that trend.

7. Conclusion and Future Work

In this work, we proposed an efficient model for threat detection on Twitter. This model consists of two main modules: ThrNet, a semantic network for modeling and representing threat knowledge, and the threat detection algorithm for tweet messages classification using ThrNet. Likewise, we showed a way for using multiple NLP techniques that analyze the content of tweet messages, such as lemmatization and NER. Also, we used machine learning algorithms like NB, DT, Logistic Regression, XGB, KNN, RF, and SVM to benchmark the performance of our proposed model DetThr using the same training and test datasets. The evaluation of the DetThr model led to good results (76% for accuracy and 75% for F1score) and showed that our model outperforms the other models.

For future work, we would like to improve the performance of the DetThr model by adding new targets into our proposed semantic network (ThrNet). For instance, a new target like careers (e.g. farmer, artist, actor, policeman, etc.) can help to increase the capability of the model to identify new types of threats which will increase, consequently, the recall of the model.

References

[1] Maaz Amjad, Noman Ashraf, Alisa Zhila, Grigori Sidorov, Arkaitz Zubiaga, and Alexander Gelbukh. Threatening language detection and target identification in urdu tweets. IEEE Access, 9:128302–128313, 2021.

[2] D. B. Alorini, D.and Rawat. Automatic spam detection on gulf dialectical arabic tweets. In 2019 International Conference on Computing, Networking and Communications (ICNC), pages 448–452, 2019.

[3] Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. Hate me, hate me not: Hate speech detection on facebook. In ITASEC, pages 86–95, 01 2017.

[4] Atta ur Rahman. Knowledge representation: A semantic network approach. In Handbook of Research on Computational Intelligence Applications in Bioinformatics, pages 55–74. IGI Global, 2016.

[5] Serhad Sarica, Jianxi Luo, and Kristin L. Wood. Technet: Technology semantic network based on patent data. Expert Systems with Applications, 142:112995, 2020.

[6] Derek L Hansen, Ben Shneiderman, Marc A Smith, and Itai Himelboim. Semantic networks. In Derek L Hansen, Ben Shneiderman, Marc A Smith, and Itai Himelboim, editors, Analyzing Social Media Networks with NodeXL, pages 115–125. Elsevier, second edi edition, 2020.

[7] Kyounghee Kwon, C. Chris Bang, Michael Egnoto, and H. Raghav Rao. Social media rumors as improvised public opinion: semantic network analyses of twitter discourses during korean saber rattling 2013. Asian Journal of Communication, 26(3):201–222, May 2016.

[8] 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, San Diego, California, June 2016. Association for Computational Linguistics.

[9] Noman Ashraf, Rabia Mustafa, Grigori Sidorov, and Alexander Gelbukh. Individual vs. group violent threats classification in online discussions. In Companion Proceedings of the Web Conference 2020, WWW '20, page 629–633, New York, NY, USA, 2020. Association for Computing Machinery.

[10] Hugo Lewi Hammer. Automatic detection of hateful comments in online discussion. In International Conference on Industrial Networks and Intelligent Systems, pages 164–173. Springer, 2017.

[11] H. L. Hammer, M. A. Riegler, L. Øvrelid, and E. Velldal. Threat: A large annotated corpus for detection of violent threats. In 2019 International Conference on Content-Based Multimedia Indexing (CBMI), pages 1–5, 2019.

[12] Khaled Bedjou, Faiçal Azouaou, and Abdelouhab Aloui. Detection of terrorist threats on twitter using svm. In Proceedings of the 3rd International Conference on Future Networks and Distributed Systems, ICFNDS '19, pages 1–5, New York, NY, USA, 2019. Association for Computing Machinery.

[13] Addie Beach. "it's so bomb": Exploring corpus-based threat detection on Twitter with discourse analysis. PhD thesis, University of Vermont, 2019.

[14] Puja Chakraborty and Md. Hanif 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), pages 1–6, 2019.

[15] Shatha Abdulaziz AlAjlan and Abdul Khader Jilani Saudagar. Machine learning approach for threat detection on social media posts containing arabic text. Evol. Intell., 14:811–822, 2021.

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

[17] Martijn Spitters, Pieter T. Eendebak, Daniël T. H. Worm, and Henri Bouma. Threat detection in tweets with trigger patterns and contextual cues. In 2014 IEEE Joint Intelligence and Security Informatics Conference, pages 216–219, 2014.

[18] Wenlin Liu, Chih-Hui Lai, and Weiai (Wayne) Xu. Tweeting about emergency: A semantic network analysis of government organizations' social media messaging during hurricane harvey. Public Relations Review, 44(5):807–819, 2018.

[19] Derek L. Hansen, Ben Shneiderman, Marc A. Smith, and Itai Himelboim. Chapter 8 - semantic networks. In Derek L. Hansen, Ben Shneiderman, Marc A. Smith, and Itai Himelboim, editors, Analyzing Social Media Networks with NodeXL (Second Edition), pages 115–125. Morgan Kaufmann, USA, second edition edition, 2020.

[20] Lu Tang, Bijie Bie, and Degui Zhi. Tweeting about measles during stages of an outbreak: A semantic network approach to the framing of an emerging infectious disease. American Journal of Infection Control, 46(12):1375–1380, dec 2018.

[21] Marya L Doerfel. What constitutes semantic network analysis? a comparison of research and methodologies. Connections, 21(2):16–26, 1998.

[22] Engels Rajangam and Chitra Annamalai. Graph models for knowledge representation and reasoning for contemporary and

emerging needs–a survey. International Journal of Information Technology and Computer Science (IJITCS), 8(2):14–22, 2016.
[23] Derek L. Hansen, Ben Shneiderman, Marc A. Smith, and Itai Himelboim. Chapter 6 - calculating and visualizing network metrics. In Derek L. Hansen, Ben Shneiderman, Marc A. Smith, and Itai Himelboim, editors, Analyzing Social Media Networks with NodeXL, pages 79–94. Morgan Kaufmann, USA, second edition edition, 2020.
[24] Ghadeer Al-Turaif and Fethi Fkih. A review on threat detection approaches in social networks. International Journal of Computer Science and Network Security (IJCSNS), 21:353–361, 2021.
[25] Joan A Sereno and Allard Jongman. Processing of english inflectional morphology. Memory & cognition, 25(4):425–437, 1997.
[26] Enrique Manjavacas, Ákos Kádár, and Mike Kestemont. Improving lemmatization of non-standard languages with joint learning. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 1493–1503, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
[27] Symeon Symeonidis, Dimitrios Effrosynidis, and Avi Arampatzis. A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. Expert Systems with Applications, 110:298–310, 2018.
[28] Deepa Yogish, TN Manjunath, and Ravindra S Hegadi. Review on natural language processing trends and techniques using nltk. In International Conference on Recent Trends in Image Processing and Pattern Recognition, pages 589–606. Springer, 2018.
[29] Òscar Garibo i Orts. Multilingual detection of hate speech against immigrants and women in Twitter at SemEval-2019 task 5: Frequency analysis interpolation for hate in speech detection. In Proceedings of the 13th International Workshop on Semantic Evaluation, pages 460–463, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.
[30] Deepa Yogish, T. N. Manjunath, and Ravindra S. Hegadi. Review on natural language processing trends and techniques using nltk. In K. C. Santosh and Ravindra S. Hegadi, editors, Recent Trends in Image Processing and Pattern Recognition, pages 589–606, Singapore, 2019. Springer Singapore.
[31] Michael Mesfin Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. Detection of suicide ideation in social media forums using deep learning. Algorithms, 13(1), 2020.
[32] Elaine J. Yuan, Miao Feng, and James A. Danowski. “Privacy” in Semantic Networks on Chinese Social Media: The Case of Sina Weibo. Journal of Communication, 63(6):1011–1031, 10 2013.
[33] Jeanette B. Ruiz and George A. Barnett. Exploring the presentation of hpv information online: A semantic network analysis of websites. Vaccine, 33(29):3354–3359, 2015.
[34] Fethi Fkih and Mohamed Nazih Omri. Learning the Size of the Sliding Window for the Collocations Extraction: a ROC-based Approach. In Proceedings of the 2012 International Conference on Artificial Intelligence: ICAI'12, Las Vegas, Nevada, USA.
[35] James A Danowski. Wordij version 3.0: Semantic network analysis software. Chicago: University of Illinois at Chicago, 2013.
[36] James A Danowski. Social media network size and semantic networks for collaboration in design. International Journal of Organisational Design and Engineering, 2(4):343–361, 2012.
[37] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. In 27th Annual Meeting of the Association for Computational Linguistics, pages 76–83, Vancouver, British Columbia, Canada, June 1989. Association for Computational Linguistics.
[38] Hailong Zhang, Wenyan Gan, and Bo Jiang. Machine learning and lexicon based methods for sentiment classification: A survey. In 2014 11th Web Information System and Application Conference, pages 262–265, 2014.
[39] Fethi Fkih and Mohamed Nazih Omri. A statistical classifier based Markov chain for complex terms filtration. In Proceedings of the International Conference on Web Informations and Technologies, ICWIT 2013, pages 175–184, Hammamet, Tunisia, 2013.
[40] Fethi Fkih, Mohamed Nazih Omri and Imen Toumia. A Linguistic Model for Terminology Extraction based Conditional Random Fields. ICCRK'2012-International Conference on Computer Related Knowledge, Sousse, Tunisia, 2013.
[41] Fethi Fkih and Mohamed Nazih Omri. “Hybridization of an Index Based on Concept Lattice with a Terminology Extraction Model for Semantic Information Retrieval Guided by WordNet”. In: Abraham, A., Haqiq, A., Alimi, A., Mezzour, G., Rokbani, N., Muda, A. (eds) Proceedings of the 16th International Conference on Hybrid Intelligent Systems (HIS 2016). HIS 2016. Advances in Intelligent Systems and Computing, vol 552. 2017. Springer, Cham.
[42] Fethi Fkih and Mohamed Nazih Omri. Information retrieval from unstructured web text document based on automatic learning of the threshold. Int. J. Inf. Retr. Res., 2(4):12–30, 2012.
[43] Sarra Ouni, Fethi Fkih and Mohamed Nazih Omri. BERT- and CNN-based TOBEAT approach for unwelcome tweets detection. Soc. Netw. Anal. Min. 12, 144 (2022). https://doi.org/10.1007/s13278-022-00970-0.
[44] Fethi Fkih and Mohamed Nazih Omri. Hidden data states-based complex terminology extraction from textual web data model. Appl. Intell., 50(6):1813–1831, 2020.

Authors’ Profiles

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.

How to cite this paper: Fethi Fkih, Ghadeer Al-Turaif, "Threat Modelling and Detection Using Semantic Network for Improving Social Media Safety", International Journal of Computer Network and Information Security(IJCNIS), Vol.15, No.1, pp.39-53, 2023. DOI:10.5815/ijcnis.2023.01.04