Journal of Engineering and Applied Sciences 13 (17): 7459-7464, 2018
ISSN: 1816-949X

Terrorist Affiliations Identifying Through Twitter Social Media Analysis Using Data Mining and Web Mapping Techniques

[1]Muhanad Abdul Elah Al-Khalisy and [2]Hashem B. Jehlol
[1]Department of Scholarships, University Information Technology and Communication, Baghdad, Iraq
[2]Electronic Computer Center, Al-Mustansiriyah University, Baghdad, Iraq

Abstract: With the increase in number of users on each day on a social media platform that generates a huge amount of data today data analysis plays a vital role. We focus on Twitter's mining role in extracting useful information that provides terrarium supporter data such as location, account name and terrarium propaganda. The proposed methods utilize Twitter streaming API to collect data, preprocessing and cleansing were performed on Tweet's data, wordlist of synonyms and antonyms words relating to terrorism get it from the dictionary, these words classified as positive and negative words. The proposed methods base on "Bag-of-Word" characteristic extraction to compute the total score of each Tweet that represents training data. Depending on the training data, the Naive Bayes classifiers classify each Tweet to positive, negative and natural. GeoJSON used to find and visualize where terrarium is located online. The results can be used by the governments and security agencies to determine relevant data to find terrarium users.

Key words: Text mining, sentiment analysis, social media, machine learning, terrarium, streaming API, GeoJSON, Naive Bayes

INTRODUCTION

In recent years, the nature of terrorism in the world has changed dramatically (Ende, 2016). The issue of terrorism is not new but recently it has become a global issue that threatens the whole world. The Federal Bureau of Investigation (FBI) defines terrorism as "The unlawful use of force and violence against persons or property to intimidate or coerce a government, the civilian population or any segment thereof in furtherance of political or social objectives"(Koplow, 2015). Twitter is one of many social media that has become a widespread and common technology for terrorism to broadcast their messages in real time all over the world. The lack of strict rules and regulations on the use of social media platforms, the ability to hide people's identity, access to very large audiences and the rapid transmission of information all these features have made social media the right choice for spreading extremist values and beliefs (Ghajar-Khosravi *et al*., 2016). This study aims to use some strategies and techniques to analyze Twitter content. Sentiment mining which focuses on calculating the general opinion of a particular group of people, extracting predefined keywords to check their frequency and perform statistical algorithms (Kocharekar and Jadhav, 2017).

Literature review: A specialized domain that applies data mining techniques over text is a "Text mining", some studies related to analysis Twitter contents generated by terrorists are descried bellow: (Kale *et al*., 2017) proposed a map reduce based Naive Bayes training algorithm, this algorithm uses only one map reduce job. Experiments are performed using Amazon EMR cluster and a large training corpus from Twitter. The results show the proposed method to be highly scalable and cost efective for larger data. The accuracy of classifer, trained using this new dataset, approached (Balamurugan and Pushpa, 2015). Proposed techniques depend on machine learning to detect the acts of terrorism more accurately. The techniques categorize the sentence into positive, negative and neutral categories. All these categories will be compared against the pervious sentence of a particular account holder based on the sentiment score for the latest and previous sentence. Machine learning is being proposed to be used in this research as it is more accurate as compared to lexicon-based approach (Ghajar-Khosravi *et al*., 2016). Examined Twitter content generated by female users who are sympathetic to the Islamic State in Iraq and Syria (ISIS). The ISIS fan girls differ in the content of their Tweets from other, non-radicalized, teenage girls and that automated text

Corresponding Author: Muhanad Abdul Elah Al-Khalisy, Department of Scholarships, University Information Technology and Communication, Baghdad, Iraq

analysis techniques can detect the differences. The basic technique proposed here is a promising step in devising techniques for quantifying the salient topics being discussed on social media platforms and should be developed further to create more fine-grained exanimations of such content.

MATERIALS AND METHODS

This study proposed method for mining text of Twitter post to detect significant information observed in huge volumes of data. It suggests the extraction of important information from terrorist supporter's data such as account name, location and ISIS propaganda. The layout of proposed method contain data gathering, JSON data cleaning, text preprocessing, sentiment classification and GeoJSON as illustrated in Fig. 1.

Data gathering: One of the most important stages for the researcher that is used to make realistic conclusion of the problem is data collection. Twitter is one of the most important social media networks. The data will be catch according to the particular key words to get posting Tweet text from Twitter in auto manner. The data that captured from Tweets have additional details such as information about sender, receiver, time, date, language, location and others. The structure of the Tweet consists of the fields mentioned above in addition to the post text. The Tweets that captured will be stored in a JSON-formatted text file where they can be better processed in the next stages (Wadhwa and Bhatia, 2014). In this research the English Language was used to mined the terrorism words that occur in the Twitter (Margono *et al*., 2014). Table represents a sample of terrorist words posted on Twitter which it used in this study.

Sample of terrorist related keywords:

- Baghdadi
- Alkhilafa
- Jihadi
- Ummah
- Kuffar
- Alraafida
- Khawarij

Twitter streaming API: The most platforms of social media now a day provides an Application Programming Interfaces (APIs) used to share data. Because of dynamic nature of Twitter API, the focus on Twitter will be largely given. The API of Twitter is made up of streaming application program interface. The streaming application program interface prepares ways for processing requests, verifying applications, manipulation imposed limits and so on. It supplies the application's client by global stream of data in JavaScript Object Notation (JOSN) format

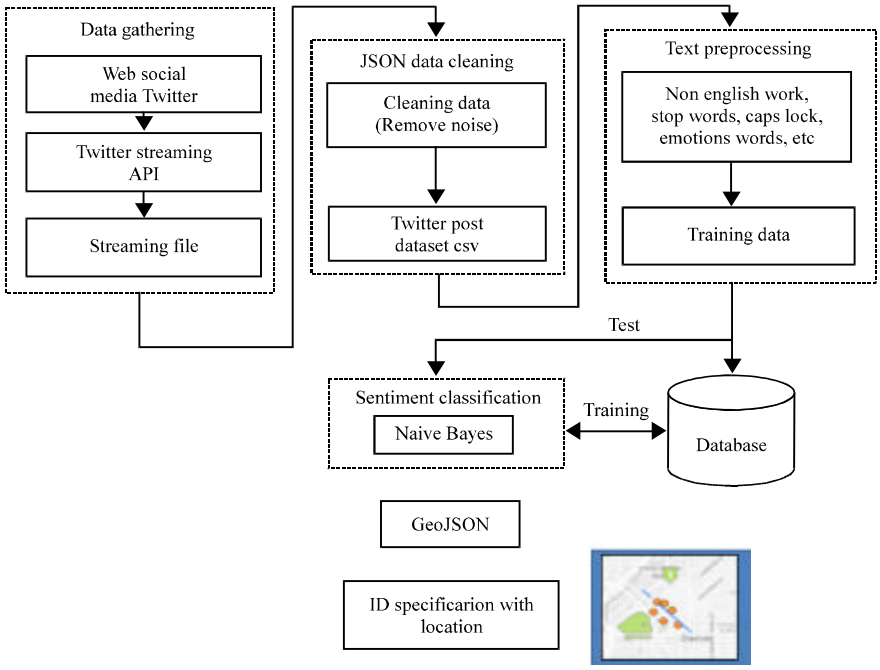
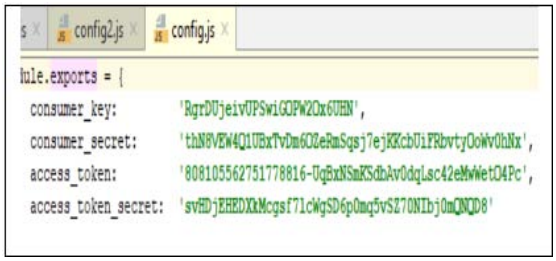


Fig. 1: The layout of proposed method

(a)



(b)

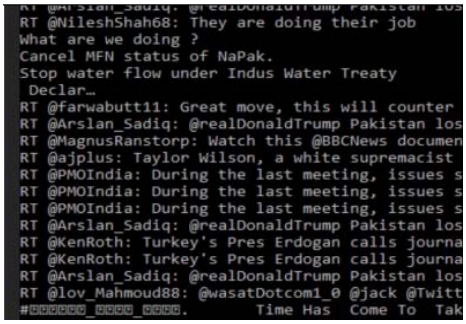


Fig. 2: The results of API where: a) Authentication keys provided to Twiter API and b) Sample streaming data that result from Twiter API

(Isah *et al*., 2014). The proposed method used a Streaming APIs for finding and fetching Tweets. For the purpose of accessing and benefiting from Twitter Streaming API, there is some information that should be obtained from Twitter application: API secret, API key, Access token secret and Access token. Figure 2a shows the authentication keys provided to Twitter API while Fig. 2b illustrates sample of streaming data that result from Twitter API.

JSON: JSON "JavaScript Object Notation" is established on a subcategory of the "JavaScript programming language" which is an uncomplicated format for individuals to reading and writing, uncomplicated for computer to analyze and procreate and lightweight data-interchange formatted. Therefore, JSON is considering ideal data-interchange language. JSON is built in the form of a set of (name/value) pairs where it can be achieved in different programming languages in the form of a record an object, a structure, a dictionary, a linked list or hash table (Isah *et al*., 2014). The application programming interfaces supplies the client application by global stream of data in JavaScript Object Notation (JOSN) format. Algorithm 1 illustrate a sample of data that result from API in JOSN format.

Algorithm 1; Sample of data resulted from API repesinted in JSON format:

profile_use_background_image: true,
profile_image_url: 'http://pbs.twimg.com/profile_images/1282407636/icon_512_normal.png'
profile_image_url_http: 'https://pbs.twimg.com/profile_images/1282407636/icon_512_normal.
prfile_banner_url: 'https://pbs.twimg.com/profile_banners/29958928/1401007898',
default_profile: false
default_profile_image: false
following: null
follow_request_sent: null
notifications: null }
geo: null
coordinates: null
place: null
contributers: null
is_quote_status: false
quote_count: 4
reply_count: 0

Dataset: A data set is a collection of related, discrete items of related data that may be accessed individually or in combination or managed as a whole entity. In the proposed method when a Tweet arrives on the stream through API, the collector parses the JSON files to extract the Tweet corresponding to specific template (ID, Tweet, Name and Coordinate) (Maynard *et al*., 2017). Tweet dataset is a brief details of dataset used in experiments. Dataset need to be cleaned and tokenized and should be in a CSV format.

Text preprocessing: Text preprocessing stage involves removal of any undesired text from tweets before classification is performed (Isah *et al*., 2014). A JavaScripts module was used to amount the sequence of fallowing tasks, all words in the upper case change to lower case, take off delimiters, clean off stop words and numbers (Ghajar-Khosravi *et al*., 2016), all letters: hash, dashes, symbols or non-English words have been removed from Tweets, any blank or redundant Tweets were deleted, any change in spellings of the transcribed Arabic expression for example the words ("Kuuffar", "Kuffar", "Kuffarr", etc.) were all substitute with "Kuffar". After that all cleansed Tweet data complain in one corpus. Table 1 shows a sample of unwanted text.

Training data: A scored wordlist generated using dictionary. This process involved searching of the synonyms and antonyms of words from dictionary that relating to terrorism. These words that related to terrorism were labeled as negative while the others words were positive and stored as CSV file format. The proposed methods base on "Bag-of-Word" characteristic extraction to computes the total score of each Tweets text by

Table 1: Sample of unwanted text and action

Unwanted content	Wanted content	Action
#word	Word	Replaced
@user_name	AT_USER	Converted
https://	SURL	Converted
More spaces‘ ‘	“ ”	Removed
Retweet	RTW	Removed
Kuffar	kuffar	Converted

Table 2: Sample of scored training data

ID	Tweet	Polarity	Values
10900	Video:cars pollution in the Arab contrary.url	Positive	2
11450	Us force bombs Islamic State terror camps; Died over 140 terrorist url	Negative	-2
11456	@keithellison If you endorse a radical group approves violent methods you probably are rad url	Neutral	0
11660	Pakistan lost more than 70 thousand real human lives in the fight against terrorism url	Negative	-1

calculating the aggregate score of words in the Tweets (Pasini and Navigli, 2017). Each word in the Tweets was matched to the words stored in the positive and negative wordlist. When there was a match the counter was incremented or decremented by a fixed number depending on the weight or value assigned to each word in the scored wordlist. This proses generates a new score for each Tweet. The resulted corpus will be used as training data for Naive Bayes classifier. Table 2 shows a sample of scored training data.

Sentiment analyzing classification: Sentiment classification covers aspect of artificial intelligence that deals with extracting the meaning from the text using various techniques. A sentiment classification can do by using machine learning or Lexicon based approaches (Kocharekar and Jadhav, 2017). This study uses machine learning approach based on classification of sentiment using training and test data sets. The technique such as “Naive Bayes” has been very successful in text classification.

GeoJSON: Web mapping library can add vector features on a map. It will be provided in multiple formats, GeoJSON, GML, GPX, etc., this method select the GeoJSON format because it is based on JSON which is light and supported by many JavaScript libraries. Usually, the post in the Twitter API contains an irrelevant metadata, the JavaScript was used to select the parts that would be used by the proposed method (Bigler *et al*., 2017). These include user ID, name, geographic coordinates, etc. For each classified Tweet, the JavaScript takes these pieces of information and creates a GeoJSON formatted file that’s contains geographic coordinates that used for detrained and maps the location of targeted ID with assigned color markers based on its polarity.

Table 3: Sample of classified data

Polarity	Tweet	Name	Coordinates
Positive	Tweet	TOI India	0.35,34.537,0
Negative	Tweet	P90sPickups	-0.35,34.199,0
Neutral	Tweet	Janet Samuels	0.375,35.17,0
Negative	Tweet	Arslan_Sadiq	-0.885,40.17,0
Positive	Tweet	must313ah	0.7775,30.17,0

RESULTS AND DISCUSSION

In this study, the proposed method shown how to use the data mining tools and other tools and program such as JavaScripts, API, JSON, etc. with social media, to detect Tweets carrying terrorist ideas and to identify the location of terrorist who owned those Tweets. The proposed performed two tasks:

Twitter data sentiment analysis: Data was collect through Twitter streaming API, selected about 10,322 Tweets with the keywords likes: (terrorism, Islamic State, Jihadi, Ummah, Kuffar and Khawarij). After that preprocessing and cleansing were performed on Tweet’s data. These data will transform to a one corpus with 10,322 Tweets. A wordlist of synonyms and antonyms words relating to terrorism get it from dictionary, these words classified as positive and negative words. The dictionary contained 1,900 positive words and 3,824 negative words. The proposed methods base on “Bag-of-Word” characteristic extraction to computes the total score of each Tweets text by calculating the aggregate score of words in the Tweets that represent training data. “Naive Bayes” classifier used which avoids the position of the word in the document. Depending on the training data, the Naive Bayes classifiers classify each Tweet to positive, negative and natural, 7,122 Tweets of them classified as negative. Table 3 shows a sample of classified data.

Sentiment mapping with JeoJSON: JavaScript with GeoJSON web mapping used for mapping sentiments on to the map. The map is produced with markers distributed across various points on it, so that, the sentiments can be easily visualized by the user. The three classe’s negative, positive and natural sentiment represented as red, green and yellow, respectively as shown in Fig. 3. This process included the coordinates which specified the location of the Tweeter. The results show that there was high percentage of negative Tweets related to terrorism compared to positive Tweets.

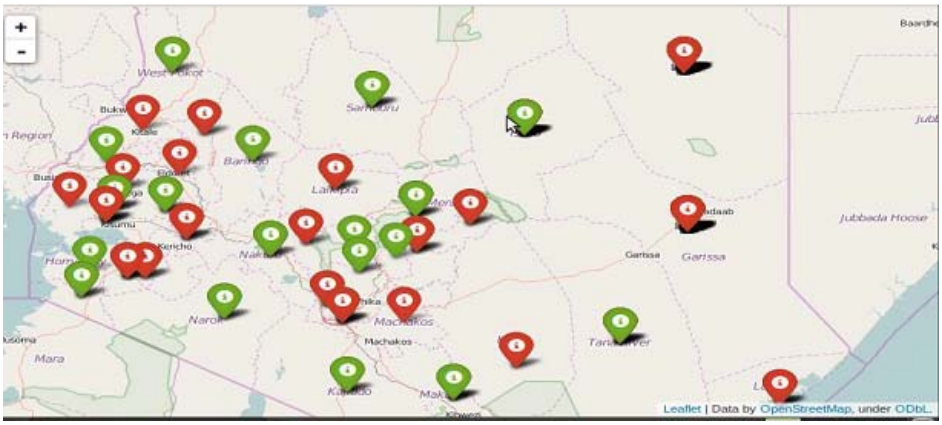


Fig. 3: The geographical representation of Tweets using GeoJSON

As a final result, this study can be used to support attempts to save the law to a large extent in countering terrorism groups as long as cleaning, processing and locating the posted data that leaved behind them on social media networks is possible.

CONCLUSION

Analysis of Twitter data helps to structure data in real time and better understand it using different techniques. The proposed methods used the Twitter APIs to access Tweets in real time. To extract the pertinent data from Tweet's raw data, text mining techniques were used, we performed sentiment analyzing to classify the Tweets as negative, positive and natural where the negative indicates a terrorism-related suspension. The results of the classification included the coordinates which specified the location of the Tweeter. The coordinates help in mapping the sentiment on to the map as visualized markers distributed across various points this greatly enhance law enforcement efforts at eradicating terrorist groups, so, long as locating.

RECOMMENDATION

We propose as the future work applying this method in other fields such as politics, arts and trade.

REFERENCES

Balamurugan, R. and D.S. Pushpa, 2015. A review on various text mining techniques and algorithms. Proceedings of the 2nd International Conference on Recent Innovations in Science, Engineering and Management, November 22, 2015, JNU Convention Center, New Delhi, India, ISBN:978-81-931039-9-9, pp: 837-848.

Bigler, A., O. Ertz, D. Rappo, S. Composto and F. Joerin *et al*., 2017. GEOPOLL-integrate cartographic questions in web forms, polls or surveys. Proceedings of the Conference on Free and Open Source Software for Geospatial (FOSS4G'17) Vol. 17, August 14-19, 2017, Seaport Hotel & World Trade Center, Boston, Massachusetts, pp: 1-9.

Ende, B.V.D., 2016. Understanding and combatting terrorist networks: Coupling social media mining with social network analysis. Proceedings of the 14th Conference on Australian Information Security Management, December 5-6, 2016, Edith Cowan University, Joondalup, Western Australia, pp: 48-51.

Ghajar-Khosravi, S., P. Kwantes, N. Derbentseva and L. Huey, 2016. Quantifying salient concepts discussed in social media content: An analysis of tweets posted by ISIS fangirls. J. Terrorism Res., 7: 79-90.

Isah, H., P. Trundle and D. Neagu, 2014. Social media analysis for product safety using text mining and sentiment analysis. Proceedings of the 14th UK Workshop on Computational Intelligence (UKCI'14), September 8-10, 2014, IEEE, Bradford, UK., ISBN:978-1-4799-5538-1, pp: 1-7.

Kale, P., R. Ranjan, T. Bagade, A. Sapkal and K. Singh, 2017. Sentiment analysis of social media data. Intl. J. Emerging Trends Technol. Comput. Sci., 6: 151-156.

Kocharekar, M. and U. Jadhav, 2017. Detecting terrorist activities using sentiment analysis in a distributed system. Intl. J. Sci. Res. Eng. Trends, 6: 285-287.

Koplow, J.L., 2015. Designation of North Korea as a state sponsor of cyberterrorism, on. SMU. Sci. Tech. L. Rev., 18: 405-405.

Margono, H., X. Yi and G.K. Raikundalia, 2014. Mining Indonesian cyber bullying patterns in social networks. Proceedings of the 37th Conference on Australasian Computer Science Vol. 147, January 20-23, 2014, Australian Computer Society, Inc, Auckland, New Zealand, ISBN: 978-1-921770-30-2, pp: 115-124.

Maynard, D., I. Roberts, M.A. Greenwood, D. Rout and K. Bontcheva, 2017. A framework for real-time semantic social media analysis. Web Semant. Sci. Serv. Agents World Wide Web, 44: 75-88.

Pasini, T. and R. Navigli, 2017. Train-O-Matic: Large-scale supervised word sense disambiguation in multiple languages without manual training data. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, September 7-11, 2017, Association for Computational Linguistics, Copenhagen, Denmark, pp: 78-88.

Wadhwa, P. and M.P.S. Bhatia, 2014. Discovering hidden networks in on-line social networks. Intl. J. Intell. Syst. Appl., 6: 44-54.