Available online at www.worldscientificnews.com

World Scientific News

An International Scientific Journal

WSN 113 (2018) 218-225 EISSN 2392-2192

Mass Violence Detection Using Data Mining Techniques

Rishabh Varma[a] and **Sartaj Ahmad**[b]

Information Technology Department, Krishna Institute of Information and Technology,
Ghaziabad - 201206, UP, India

[a,b]E-mail address: rishabhvarma22@gmail.com , sartajahmad2u@gmail.com

ABSTRACT

The world is now witnessing a tectonic shift in the way in which people react to social and economic impacts such as rise in fossil fuel prices, implication of new rules and regulations, and other situations which directly affect the emotions of a certain group of people. Violence is the most widely used way of expressing anger and discontent for a particular situation which might have occurred. Such actions can cause loss of millions of dollars and precious lives of people who come in way of such protests. These protests are mainly conducted through social media platforms such as twitter as it is not possible to personally communicate to tens of thousand people to accumulate at a certain place, therefore it is extremely important as well as necessary to keep an eye on the social media statuses and updates of people in the times of crisis and heavy tension. This paper aims to collect the tweets of people uploaded on twitter and then process them to find out the location, time and intensity of the mass violence so that the responsible authorities can handle the situation and prevent violence.

Keywords: Data mining, Text mining, Predictive model, Tweet analysis

1. INTRODUCTION

In the times of crisis when people are outraged because of the happening of a certain event which directly effects the life of a certain group of people in a certain way it is extremely

(Received 02 October 2018; Accepted 26 October 2018; Date of Publication 31 October 2018)

important to keep an eye on the social media activities of people as the social media is the most used method of communication throughout the world and has become an integral part of lives of people in every possible way. The protests can take a violent face and can cause loss of lives and capital in no time, such happenings can be foreseen and prevented by using data mining techniques and tracking the activities of people in response to the occurrence of a certain event or activity [1].

This paper focuses on collecting the tweets containing certain key-words and hash-tags. Hash-tags are easy to interpret and are best way to search a tweet in the enormous heap of tweets. The hash tags are related to each other in the form of graphs [3], hash-tags graphs are highly effective in forming relations between the tweets and the keywords in the database. The collected tweets may be written in any language of the world therefore all the collected tweets are converted into English by using google translator. Once the required tweets are collected and translated then certain text-preprocessing techniques are used to convert the tweets in a form over which the classification and clustering techniques can be applied [2]. The text-preprocessing techniques involve *tokenization, data-standardization, emoticon-conversion, stemming* and *abbreviation analysis* [13]. The further preprocessing of data involves techniques such as n-gram and tf-idf, these algorithms make the tweets easier to interpret and apply data mining techniques. Then AGF (Associativity Gravity Force) is used to classify tweets in predefined topics using Kr and CIMAWA [19]. Finally the classification algorithms do the work and classify [9] the tweet based on certain predefined parameters such as location, time & date, number of people and weapons that may be used for violence [6].

The given model can efficiently find out the location of the mass violence as well as the time, date, number of participants and the intensity of violence of multiple mass violence conditions at a given time [11]. Using hash-tags make it easier to classify the tweets and form the relations based on the hashtag graph [5]. Also, the emoticons used will be converted to words assigned to them based on the emoticons. Therefore the model proves to be highly efficient and useful in multiple situations involving crisis.

2. PROPOSED MODEL

Hashtag graph [3]

A general hashtag graph H = {X,E}, where E represents the edge of the link between two hashtags. Each edge emn forms an undirected link between two hashtags hm and hn. The hashtags present in the tweets are related to each other if they occur atleast once in a tweet. The hashtags are of three types 1) first category contains hashtags which are closely and directly connected to the subject (eg. “#hike_in_tax”), 2) second are those which show sentimental relation with the main subject (eg. “#time_for_rage” and “#make_them_suffer”), 3) the third category contains those tweets which indicate the sentiment polarity on the subject (eg. “#no_violence” and “#need_redemption”). The graph is used to form relationships between tweets and show sentiment polarity associated with each of them [21].

- **Removing junk**

Those tweets which do not covey any meaning in human readable language are considered “junk” and are deleted from the database. All the other tweets are considered “useful” and are further used data processing.

For example – any tweet which is replicated from the same address shall be considered junk. Any tweet containing humanly unreadable language shall be considered junk.

- **Relating tweets**

The useful tweets once collected will be given a certain unique integer value corresponding to every tweet in the database for further ease of processing later on.
For example – the first tweet encountered shall be assigned a serial number 1, the second tweet encountered shall be given number 2 and so on, so that we can identify the tweets at the time of classification of tweets.

- **Text-preprocessing [2]**

Text preprocessing involves various steps involved to convert the text into a structured form which can be easily processed by data mining algorithms [23].

T*okenization*: - Every Tweet is split into meaningful words called tokens. Example - "Morning walk is a bliss" is converted to "Leader" "is" "a" "traitor".

***Data standardization*: -** It involves converting all words in the tweet in standard form, converting all words in lower case [24]. Example - "The petrol prices are hiked again by 10%" is converted to "the petrol prices are hiked again by 10%".

***Emoticons conversion:* -** The emoticons present in the text messages are assigned a keyword based on the expression they convey [16].

[16] The emoticons are classified into following two categories: -

Positive emoticons- these are the emoticons which convey positive sentiment and are replaced by positive words based on the symbol.

Negative emoticons - these emoticons reflect the sad or disturbed sentiments of the subject and are thus replaced by negative words.

***Stop-word-removal*: -** All the words in the tweets which do not convey a special meaning are removed like a, the, then, etc. [15].

***Stemming*: -** It involves obtaining the root word corresponding to every word by dropping suffixes ling -ing, -ion, etc. [7, 14].

***Abbreviation analysis*: -** the abbreviations present in the tweets are replaced by their full forms. Example GOI by government of india, ONGC by Oil and Natural Gas Corporation Limited, etc.

- **AGF (Associative Gravity Force)**

After evaluating the frequent patterns using tf-idf, the model can also evaluate the AGF(Associative Gravity Force) values between each pattern pairs with the help of two other parameters namely Kr(Keyword rating) and CIMAWA(Concept for the Imitation of the Mental Ability of Word Association).

- **Keyword rating**

Kr of each frequent word, which is obtained from tf-idf. f(w) represents the number of occurrences of w, in one document. Word with the highest value of Kr will be taken as the most important one [20].

- **CIMAWA [19]**

Instead of evaluating CIMAWA value for each pair of frequent patterns, x and y, CIMAWA (x(y)) is evaluated only if y occurs after x in any of the tweet. Co-occurrence (Cooc (x, y)) is the number of tweets in which both x and y occurred together. f(y) represents the number of tweets in which y occurred, in other words, x in the case of f(x). Ꟶ is a damping factor, whose value is defined in the interval 0 and 1. Through various case studies it has been proved that the best value for Ꟶ is 0.5.

Kr(x) and Kr(y) evaluate the importance of x and y respectively in one document, and CIMAWA(x(y)) evaluates.

- **Evaluate AGF[18]**

The probability of co-occurrence of x and y together. Then next step is the evaluation of Associative Gravity Force between x and y by using the equation defined in (5) 1. AGF evaluates the attraction between x and y. If the AGF (x(y)) is large, that means, the attraction between x and y is very high, i.e., the chances of occurrences of x and y together is very high. So we cluster the words (frequent patterns) using these AGF values.

- **Cluster Patterns using AGF**

A new method for detecting multitopic structures in text documents, called Associative Gravity [18], is based on a text-mining method entitled CIMAWA, which imitates the human ability of word association. Specifically, Associative Gravity utilizes word association to detect different topics in a text. The authors named it Associative Gravity because of its resemblance to the physical law of gravitation, that is, mass and attraction. The mass corresponds to the importance of words in a text and the attraction to the asymmetrical associative word space. The innovative characteristic of the described topic detection method is supplied with asymmetrical associative word space provided by CIMAWA. A comparative case study proves the capability of Associative Gravity to separate different topics at very high accuracy.

- **Support Vector Machines [4,10]**

The resulting stream of words after the text pre-processing step are processed by SVM Algorithm in order to classify the messages on the basis of the value of the respective parameters. SVM's are supervised learning models which are used for classification and regression analysis of data used [8]. A SVM model represents examples as points in space, different classes of examples are divided by a certain gap which must be as wide as possible. New examples when mapped into the space are predicted to belong to a class of examples based on which side of the gap they fall.

Predicting Location

The location can be predicted using svm techniques on the keywords achieved obtained from the tweet, if the tweet is threatening then the coordinates of those places will be marked which are discussed in the tweet, as they may prove to be a target place for accumulation of people. In case the place is not mentioned in the tweet then the latitude and longitude of the sender of the tweet will be marked on the map. The graph of the possible location is based on the coordinates achieved from the latitudes and longitudes of all the landmarks on earth. Once the graph is plotted then the places which have most number of markings can be easily identified as the place where violence may commit. Y-axis gives information about the north and south directions whereas the X-axis gives information about the east and west directions.

For example – a tweet that says "the world shall perish, we must take the revenge of the fallen one at Dilshan garden tomorrow #revenge_for_brothers". In this case the location of Dilshan garden shall be marked on the graph.

Predicting time & date

The time and date can be taken from that mentioned in the tweet. In case day of the week is given then it is converted to respective date. The time & date graph contains date on the Y-axis and time on the X-axis. The value of time lies between 12:00:00 AM to 11:59:59 PM and the value of date lies between the present date and a week ahead, ie current date to current date + 7. Based on the plotting of the graph we can estimate the time and date of the violent act.

For example – "tomorrow at 7 pm in front of Jantar Mantar #fight_for_nation". In this case the time will be marked on a separate graph for the process of classification.

Predicting number of protestors & ammunition

The number of protestors are calculated based on the number of tweets having those keywords, the number of people liked or reacted to the tweet, number of comments in favour of the tweet and number of retweets for a given tweet. The popularity of a person directly effects the weight of the tweet as the number of likes and re-tweets are significant. We will also be looking for words which resemble any kind of ammunition from our database having list of words which have a list of ammunitions. It will also consider extra aggressive words, these aggressive words will be considered while giving weight to tweets when ammunition is considered. In the graph, the Y-axis will represent number of people and the X-axis will represent the intensity of ammunition. The ammunitions will be weighted in the database on the basis of the amount of damage they can do.

For example - "Blood shall be shed with the boom of the guns #revenge". In this case the model will focus on the aggressiveness of the words in use and also on the number of times the tweet is liked and re-tweeted. It will assign the weight by adding all such parameters which show the popularity of the tweet.

3. CONCLUSIONS

The final result is based on the number of tweets belonging to a particular area of the graph, the relation between the graphs is obtained through the fact that we have already provided a unique integer value to every tweet. Finally we can know which places are more prone to attacks, at what time and date the attack will take place, how many people will participate in the protest and what will be the intensity of attacks [12]. This much information is sufficient for the peace authorities to stop such protests on time by various tactics. Therefore the proposed model is efficient enough to process tweets and help in stopping the wastage of millions of dollars and lives of innocent people. Text-preprocessing makes the text easier to be interpreted by the data mining algorithms, the n-gram method increases efficiency, similarly tf-idf, AGF are used to categorize text based on the predefined topics and then we apply SVM [10] on the tweets based on different set of parameters in order to find out the location, date & time, number of people and intensity of the protest.

4. FUTURE SCOPE AND APPLICATIONS

The proposed model can be used in situations where sentiment analysis is required to achieve the desired result and use it for various different purposes such as criric reviews for hotels [17] movies, videos, etc. Sentiment analysis methods till now have been used to detect the polarity in the thoughts and opinions of all the users that access social media. Businesses are very interested to understand the thoughts of people and how they are responding to all the products and services around them [22]. Companies use sentiment analysis to evaluate their advertisement campaigns and to improve their products. Companies aim to use such sentiment analysis tools in the areas of customer feedback, marketing, CRM, and e-commerce.

References

[1] Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar. Introduction to data mining. 2006 Pearson Addison-Wesley.

[2] C. Paper, ―Preprocessing Techniques for Text Mining Preprocessing Techniques for Text Mining. *J. Emerg. Technol. Web Intell. no.* October 2014.

[3] Wang, Yuan, et al. Hashtag graph based topic model for tweet mining. Data Mining (ICDM), 2014 IEEE International Conference on. IEEE.

[4] Christopher JC Burges. 1998. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery* 2, 2 (1998), 121–167.

[5] Pak, Alexander, and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. *LREc.* Vol. 10. No. 2010. 2010.

[6] Jiawei Han, Micheline Kamber, and Jian Pei. 2006. Data mining: concepts and techniques. Morgan Kaufmann.

[7] David A Hull et al. 1996. Stemming algorithms: A case study for detailed evaluation. *JASIS* 47, 1 (1996), 70–84.

[8] Hideki Isozaki and Hideto Kazawa. 2002. Efficient support vector classifiers for named entity recognition. In Proceedings of the 19th International Conference on Computational linguistics Volume 1. Association for Computational Linguistics,

[9] Mike James. 1985. Classification algorithms. Wiley - Interscience.

[10] Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. Springer.

[11] Mehmed Kantardzic. 2011. Data mining: concepts, models, methods, and algorithms. John Wiley & Sons.

[12] Leah S Larkey and W Bruce Croft. 1996. Combining classifiers in text catego-rization. In Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 289–297.

[13] Liddy, E.D. 2001. Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc.

[14] Julie B Lovins. 1968. Development of a stemming algorithm. MIT Information Processing Group, Electronic Systems Laboratory.

[15] Catarina Silva and Bernardete Ribeiro. 2003. The importance of stop word removal on recall values in text categorization. In Neural Networks, 2003. Proceedings of the International Joint Conference on, Vol. 3. IEEE, 1661–1666.

[16] Ahmad, Sartaj & Varma, Rishabh. (2018). Information extraction from text messages using data mining techniques. *Malaya Journal of Matematik.* S. 26-29. 10.26637/MJM0S01/05.

[17] Jin, Lianjing, et al. A Text Classifier of English Movie Reviews Based on Information Gain. Applied Computing and Information Technology/2nd International Conference on Computational Science and Intelligence (ACIT-CSI), 2015 3rd International Conference on. IEEE, 2015.

[18] Klahold, Andre, et al. Using word association to detect multitopic structures in text documents. *IEEE Intelligent Systems* 29.5 (2014): 40-46.

[19] Ansari, Fazel, Patrick Uhr, and Madjid Fathi. Textual meta-analysis of maintenance management's knowledge assets. *International Journal of Services, Economics and Management* 6.1 (2014): 14-37.

[20] Wang, Hongning, Yue Lu, and Chengxiang Zhai. Latent aspect rating analysis on review text data: a rating regression approach. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACm, 2010.

[21] Ma, Zongyang, Aixin Sun, and Gao Cong. On predicting the popularity of newly emerging hashtags in twitter. *Journal of the Association for Information Science and Technology* 64.7 (2013): 1399-1410.

[22] Ramya, R. S., et al. Feature Extraction and Duplicate Detection for Text Mining: A Survey. *Global Journal of Computer Science and Technology* 16.5 (2017).

[23] Feldman, Ronen. Techniques and applications for sentiment analysis. *Communications of the ACM* 56.4 (2013): 82-89.

[24] Mohamad, Ismail Bin, and Dauda Usman. Standardization and its effects on K-means clustering algorithm. *Research Journal of Applied Sciences, Engineering and Technology* 6.17 (2013): 3299-3303.

[25] Thompson, Dominic, and Ruth Filik. Sarcasm in written communication: Emoticons are efficient markers of intention. *Journal of Computer-Mediated Communication* 21.2 (2016): 105-120.

[26] Jibril, Tanimu Ahmed, and Mardziah Hayati Abdullah. Relevance of emoticons in computer-mediated communication contexts: An overview. *Asian Social Science* 9.4 (2013): 201.