IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CRIME AND FAKE NEWS DETECTION IN SOCIAL MEDIA USING MACHINE LEARNING

[1] Kovendan.V, [2]NanthaKumar.B, [3]Karthikeyan.S, [4]Subash.K, [5]Tharik Anwar.S
[1]Assistant professor [2,3,4,5] Bachelor of Engineering, Final Year CSE
Department of Computer Science and Engineering,
Arasu Engineering College, Kumbakonam, Tamil Nadu, India,

ABSTRACT

Social media is one of the greatest source, which creates high impact on grouping the public easily. So these free sources has the possibilities of making roomers and fake news to spoil an individual and leads the crime in a society. Crime is defined as an act harmful not only to the individual involved, but also to the community as a whole. Crimes are social nuisances that place heavy financial burdens on society. Here we look at use of data mining followed by sentiment analysis on online social networks, to help detect the crime patterns. Twitter is an online social networking and micro blogging service that enables users to post brief text updates, also referred to as "tweets". These updates can convey important information about the author. A filter was designed to extract tweets from cities deemed to be either the most dangerous or the safest in the United States (US). A geographic analysis revealed a correlation between these tweets and the crimes that occurred in the corresponding cities. Over 100,000 crime-related tweets were collected over a period of 20 days. Sentiment analysis techniques were conducted on these tweets to analyze the crime intensity of a particular location. This type of study will help reveal the crime rate of a location in real-time. Although the results of this test helped in detecting crime patterns, the sentiment analysis techniques did not always guarantee the proper results. We can conclude with applications of this type of study and how it can be improved by applying media to text processing techniques and also added to the system that are detect the current location of user.

Keywords: Crime Detection, SVM Classification, Fake Message Detection, Sentimental Analysis

I. INTRODUCTION

National security concern is the primary goal of any nation. Criminology studies focus on identifying criminal characteristics. The application of data mining techniques can help with this identification. Crime analysis, a part of criminology, is a law enforcement function that involves the systematic analysis of identifying and analyzing both patterns and trends in crime and disorder. In the current world, the criminals are becoming technologically sophisticated, often expressing their emotions on the web. The World Wide Web's phenomenal growth has resulted in more users expressing their opinions online. Customers use these opinions to buy a product, conduct market analysis, and so forth. This work was conducted in an attempt to accomplish to conduct geographic analysis of social media within selected cities, Analyze certain city intensity by applying sentiment analysis techniques to collected tweets. Identify the applications needed for this type of study.

1.1 Contributions

This work was proposed to better understand the crime intensity of a particular location, in almost real-time, through the online social media. As stated earlier in this paper, the existing studies draw the crime patterns using the historic data which lacks the real-time feasibility. As technology is growing rapidly, the data exchange can be done at a glance. Using the power of online social media, we believe this approach could be very useful in drawing patterns for crime detection. The approach used in this study began with the identification of the top ten crime prone cities and the top ten safest cities in the United States as determined by Forbes [3, 4]. The tweets generated within certain geographical area around these cities were then collected. The data collection process ran for nearly 21 days; which resulted in over 100,000 tweets in our database. Geographic analysis is performed using the density of population in the respective cities. The results drawn from this phase matches the pattern mentioned in the Forbes articles. Sentiment analysis was applied over the collected crime-related tweets to measure the crime intensity of a particular location. Both Stanford's Recursive Deep model [5] and the dictionary-based approach, using Affective Norms for English words (ANEW) [6, 7], were used to conduct the sentiment analysis technique. The sentiment obtained from these techniques was used to identify a location's intensity in almost real-time. As machines become increasingly capable, tasks considered to require "intelligence" are often removed from the definition of AI, a phenomenon known as the AI effect. For instance, optical character recognition is frequently excluded from things considered to be AI, having become a routine technology.[5] Modern machine capabilities generally classified as AI include successfully understanding human speech, competing at the highest level in strategic game systems (such as chess and Go), autonomously operating cars, intelligent routing in content delivery networks, and military simulations.

II. PROPOSED SYSTEM

The main challenge behind crime data mining is to understand patterns in criminal behavior in order to predict crime and prevention. Any research that can assist in solving crimes is preferred to protect individuals. A number of studies examined data obtained from either a sheriff's office or a Crime Analysis Unit. Clustering and Series Finder algorithms, respectively, were applied to the data in an effort to predict crime. Twitter, a powerful online social network, was used in this study to detect crime in almost real-time. The top ten most dangerous cities in the US, as listed by Forbes magazine, were chosen for examination; the top ten safest cities were also examined for comparison

III. ARCHITECTURE DIAGRAM:

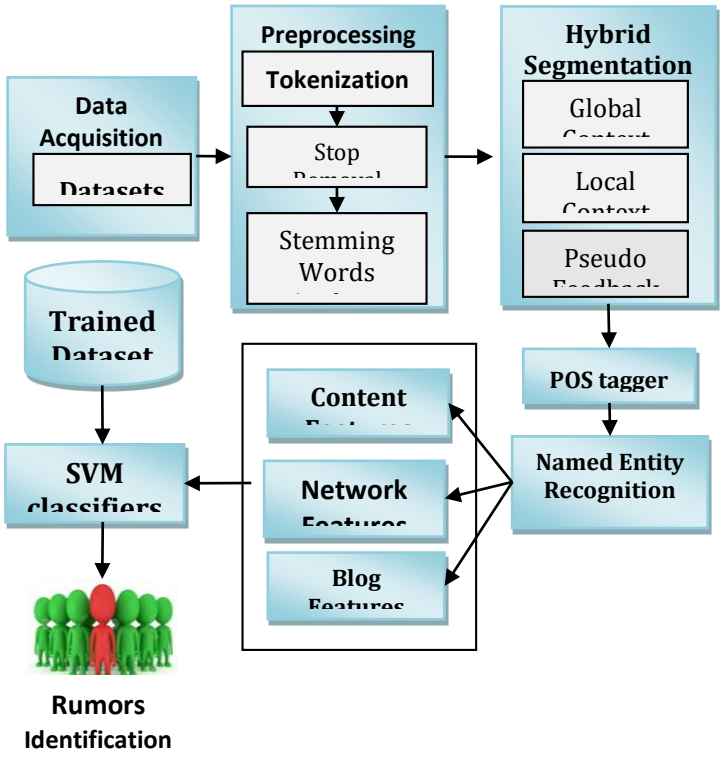

Figure 3.1 System Architecture of Crime and Fake Detection

IV. Modules of Crime Detection

4.1 Data Acquisition:

Facebook is an online social networking service that enables users to send and read messages, images as well as videos posts, . Registered users can read and post, but unregistered users can only read them it is also only if the concerned data owner provides permission, then only it is possible.

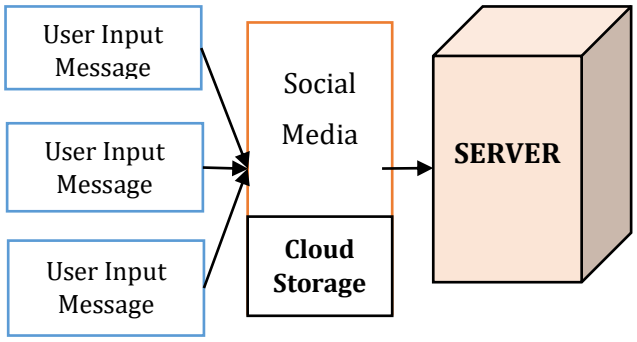

Figure 4.1 Data Acquisition from User to Cloud

Users access Facebook through the website interface or mobile device app. In order to have an opinion about the user, his posts have to be examined. Therefore, using Facebook API, all posts posted by user are crawled first. In this study, we tried to examine the user with not only his posts but also his friends' posts. However, crawling all friends' posts is a huge overload, and misleading since Facebook following mechanism does not show an actual interest every time. People sometimes tend to follow some users for a temporary occasion and then forget to un-follow. Sometimes they follow some users just to be informed of, although they are not actually interested in. There are also friends that do not post for a long time, but still followed by the user. In this module, we can upload the datasets as CSV file. It contains following id, followers id, time stamp, user following, user followers and posts. The data of entire consumers of the Facebook has been examined and their entities are analyzed for better process. The data that has been acquired are the posts that has been done by the users, messages that has been sent and received and so on it continuous.

4.2 Preprocessing:

For named entities to be extracted successfully, the informal writing style in posts has to be handled. Before real data has entered our lives, studies on the area were being conducted on formal texts such as news articles. Generally named entities are assumed as words written in uppercase or mixed case phrases where uppercased letters are at the beginning and ending, and almost all of the studies bases on this assumption. However, capitalization is not a strong indicator in posts like informal texts, sometimes even misleading. As the example of capitalization shows, the approaches have to be changed. To extract named entities in posts, the effect of the informality of the posts has to be minimized as possible.

4.3 Hybrid segmentation:

The well preserved linguistic features in these posts facilitate named entity recognition with high accuracy. Each named entity is a valid segment. The method utilizing local linguistic features is denoted by HybridSeg NER. It obtains confident segments based on the voting results of multiple off-the-shelf NER tools. Another method utilizing local collocation knowledge, denoted by HybridSeg NGram, is proposed based on the observation that many posts published within a short time period are about the same topic. HybridSeg NGram segments the posts by estimating the term-dependency within a batch of posts. The segments recognized based on local context with high confidence serve as good feedback to extract more meaningful segments. The learning from pseudo feedback is conducted iteratively and the method implementing the iterative learning is named HybridSegIter.

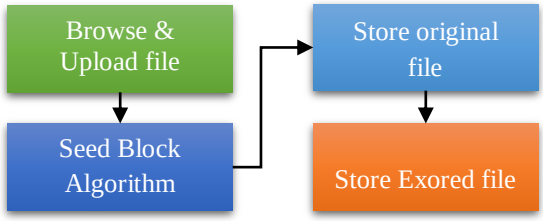

Figure 4.2 Hybrid segmentation for storing files

4.4 Named Entity Recognition:

Named Entity Recognition can be basically defined as identifying and categorizing certain type of data (i.e. person, location, organization names, and date-time and numeric expressions) in a certain type of text. On the other hand, posts are characteristically short and noisy. Given the length of a posts, and restriction free writing style, named entity recognition on this type of data become challenging. After basic segmentation, a great number of named entities in the text, such as personal names, location names and organization names, are not yet segmented and recognized properly. Part of speech tagging is applicable to a wide range of NLP tasks including named entity segmentation and information extraction. Named Entity Recognition strategies vary on basically three factors: Language, textual genre and domain, and entity type. Language is very important because language characteristics affect approaches. Assign each word to its most frequent tag and assign each Out of Vocabulary (OOV) word the most common POS tag.

4.5 Performance Evaluation

In this module, we can evaluate the process of the system using accuracy rate and normalized utility. Our proposed system provides improved accuracy rate and normalized utility. Once the messages that has been received by the receiver form the sender side, it analysis the data for negative words and positive words and incase of presence of negative it warns the user and if the process continuous more than 3 or 4 times it will make a suggestion and blocks the users who are associated with the negative contents. And also the posts that are shared by the data owner can be shared as a public and private. In public the posts can be viewed by the entire users present in the data owner's profile, whereas in private mode only the owner permitted users can access the posts that

V. CONCLUSION

A crime pattern can be detected, nearly in real-time, when online social media is monitored. The proposed system is used to identify the crime intensity of a specific location and more accurate results can be drawn from social media. Outcomes from geographic data analysis accompanied on different tweets providing a perfect picture of the criminal trends in several different places. The crime strength day-wise positively correlated with crime statistics from cops, which ultimately prove the assumption. To be more clear-cut, we analyzed the specific twitter accounts which tweet only about the crime scenarios take place in a place based on sheriff data and visualized. The results gathered from this study were positive. An advanced technology for analyzing the sentiment will aid in segregating a threatening murderer from tweets within a specific location. Video-to-text processing, image-to-text processing, and data from various online sources would also help to improve the data accuracy. This system supports informing others of the crime pattern both within and around their location, ultimately assisting them with staying in a safe zone and helps the social media to improve the accuracy for monitor and detect the crime. In future we can extend the work towards alerting the users through online to avoid the threatening and crimes.

REFERENCE

[1] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2003.

[2] J. Goldenberg, B. Libai, and E. Muller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth," Marketing letters, vol. 12, no. 3, pp. 211–223, 2001

[3] H. W. Hethcote, "The mathematics of infectious diseases," SIAM review, vol. 42, no. 4, pp. 599–653, 2000.

[4] M. E. J. Newman, "The structure and function of complex networks," SIAM review, vol. 45, no. 2, pp. 167–256, 2003.

[5] S. Wang, X. Hu, P. S. Yu, and Z. Li, "MMRate: inferring multi-aspect diffusion networks with multi-pattern cascades," in Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2014, pp. 1246–1255.

[6] Y. Tang, Y. Shi, and X. Xiao, "Influence maximization in near-linear time: a martingale approach," in Proc. ACM Int. Conf. Special Interest Group Manage. Data, 2015.

[7] L. Weng, A. Flammini, A. Vespignani, and F. Menczer, "Competition among memes in a world with limited attention," Scientific Reports, vol. 2, 2012.

[8] S. A. Myers and J. Leskovec, "Clash of the contagions: Cooperation and competition in information diffusion," in Proc. 12th IEEE Int. Conf. Data Mining, 2012.

[9] V. Kovendan "Novel Approach for Reliable Communication in Wireless Sensor Networks Using DDRA" ISSN: 2455-2631 Volume 4, Issue 5, © May 2019 IJSDR

[10] Y. Bi, W. Wu, and Y. Zhu, "Csi: Charged system influence model for human behavior prediction," in Proc. 13th IEEE Int. Conf. Data Mining, 2013.

[11] M. Coscia, "Competition and success in the meme pool: a case study on quickmeme. com," in Proc. Int. AAAI Conf. Weblogs Soc. Media, 2013.

[12] I. Valera and M. Gomez-Rodriguez, "Modeling adoption and usage of competing products," in Proc. 15th IEEE Int. Conf. Data Mining, 2015.

[13] V. Kovendan, S. Karthik "Data Center Workload Management of Scheduling, Resource Processing" Volume 1 Issue 3 HBRP Publication 2018.

[14] B. A. Prakash, A. Beutel, R. Rosenfeld, and C. Faloutsos, "Winner takes all: competing viruses or ideas on fair-play networks," in Proc. 21rd Int. Conf. World Wide Web, 2012.

[15] B. Karrer and M. E. J. Newman, "Competing epidemics on complex networks," Physical Review E, vol. 84(3): 036106, 2011.

[16] X. Rong and Q. Mei, "Diffusion of innovations revisited: from social network to innovation network," in Proc. 22th ACM Int. Conf. Inf. Knowl. Manage.,2013.

[17] N. Pathak, A. Banerjee, and J. Srivastava, "A generalized linear threshold model for multiple cascades," in Proc. 10th IEEE Int. Conf. Data Mining, 2010.