Suspicious Activity Detection of Twitter and Facebook using Sentimental Analysis

Saeed Al Mansoori [1], Afrah Almansoori [2], Mohammed Alshamsi [3], Said A. Salloum [4], Khaled Shaalan [5]

[1]Applications Development and Analysis Section, Mohammed Bin Rashid Space Centre, Dubai,United Arab Emirates
[2]General Department of Forensic Science and Criminology, Dubai Police GHQ, Dubai,United Arab Emirates
[3]Information Technology Department, Ministry of Health and Prevention, United Arab Emirates
[4]Research Institute of Sciences & Engineering, University of Sharjah, Sharjah, United Arab Emirates
[5]Faculty of Engineering & IT, The British University in Dubai, United Arab Emirates

Abstract – **The purpose of this study is to evaluate the criminal behavior on the social media platforms and to classify the gathered data effectively as negative, positive, or neutral in order to identify a suspect. In this study, data was collected from two platforms, Twitter and Facebook, resulting in the creation of two datasets. The following findings have been pointed out from this study: Initially, VADER twitter sentimental analysis showed that out of 5000 tweets 50.8% people shared a neutral opinion, 39.2% shared negative opinion and only 9.9% showed positive opinion. Secondly, on Facebook, the majority of people showed a neutral response which is 55.6%, 38.9% shared positive response and only 5.6% shared negative opinion. Thirdly, the score of sentiments and engagement in every post affects the intensities of sentiments.**

Keywords – **Criminal behavior, social media platforms, Twitter, Facebook, Part-of-Speech tagging, Valance Aware Dictionary.**

1. Introduction

Modern technologies are riveting phenomenally across immediate communication in the age of information technology. Digital and interactive technologies affect profoundly human behavior and social experiences. Digital data expand rapidly in the era of knowledge and internet. Research work shows that various social networking platforms such as Facebook, Twitter, Instagram, etc. are increasingly being utilized for the dispersal of people's ideas, perspectives and values. Many reasons can be considered as motives for using social media platforms such as business communication, news updates, job advertising, video and social gaming and many more. However, it has also intensified the abuse of social networking platforms, which contributes to a rise in criminal activities by distributing violence, sharing abuse, threatening others, and thus increasing questionable internet content.

People communicate with their families and friends through social media. In the beginning, they assume that these social media platforms are used to communicate with others when time transcends consumers' involvement through the use of those platforms, so social networking becomes a crucial aspect of their lives. Most likely, they log in to social media networking platforms at least once a day, thus it becomes a daily routine task. The two top used social media platforms due to their immersive functionalities are Facebook and Twitter. Their benefits are numerous as they include easy and fast contact and communication; prompt response, a wide range of search options, a comfortable and engaging environment, identity management, graphical user interface, and much more. In contrast, some aspects can manipulate those platforms' negativities and vulnerabilities.

1.1. Motivation

In addition to the advantages and usages of social media networking platforms, several other issues can manipulate the detriments and expose them. Since it is based on human experiences in an individual environment, anyone will sign up for these platforms to use their thoughts and desires; hence it's being shared with others. Every nasty person may use this opportunity to disrupt behaviors like suspicious chatting because online messaging is a very common feature of almost any social media networking platform, or someone might share inappropriate actions with others that could affect others' behaviors. Therefore, in order to identify any suspicious behavior on the basis of social media messages, it is possible to filter information and accounts using a sentimental analysis system. Essentially, the system of sentimental analysis examines and categorizes the data into three classes that are positive, negative, and neutral. Such a system can be used to detect suspicious activity on social media networking or custom social platforms.

1.2. Social Media System Architecture

The social media system architecture is built on LAMP (Linux, Apache, MySQL, PHP/Python/Perl) stacks. It is divided into two user spaces, collaborative value ecosystem (CVE) and leadership value ecosystem (LVE). These traditional ecosystems are composed of the user group, the service platform, and the contents provider [1]. The rise in the user diversity parameter is a growth indicator. For service platforms to attract diverse segments, investing equal resources in both user spaces is recommended. Equilibrium requires harnessing the strength of each space to fortify the other. Feedback from the information-generating capability of the CVE can be harnessed to improve the service level in the LVE optimizing both their service delivery processes. The primary access point has shifted from user browsers to autonomous application platforms. A dynamic architecture that remains consistent across devices is strategic to optimize user experience (USX), USX. It entails ease of navigating the site and its aesthetic design. Recognizing that the user group does not always double as the content provider is important. Specialized tools to optimize value creation capabilities are created with the optimal behavior of intensive users, such as gamers being the lower benchmark. Data analytics enables continual improvement of site features by delineating, the most popular ones, and the least used.

In the collaborative value ecosystem, direct user feedback besides the patterns identified by data analytics informs site improvements. Provisions that allow advanced users such as developers to tweak the front-end capabilities are available [2]. Application programming interfaces are available to developers to inspire them to create applications for the system. These are essential for improving the site's functionality as Figure 1 show the architecture of social media networking platform [3]. Access to the back-end is limited to service platform engineers; most social networking sites are secretive about their code and web frameworks [2]. Ethical hackers can exploit bugs in the system to access the source code. The practice is crucial in optimizing cyber security.

Social networks' value is hinged on its function of optimizing real-time communication. This utility is enabled through load balancing modules that buffer user queries, status updates, and other content uploads before they are transmitted to the data servers [1]. The request management capability is optimized by a preemptive approach, and the engineers recognize that while streamlining internal functioning is essential, preparing for the events of data failure due to external disturbances is equally important.
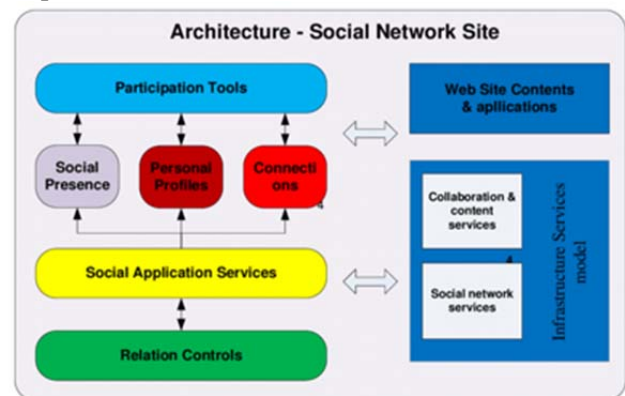

Figure 1. Architecture of Social Media networking platform [3]

2. Literature Review

In [4] studied the indication that online discussion sites have been compromised with promotional chat. However, it is not clear that these knowledge-sharing sites in which consumers discuss and inform about products provide a useful platform for advertisement manipulators. Also, it analyzes the extent and the effect of these coercive interventions in the online review world. The authors present an easy and efficient method of figuring out how reviews are manipulated. The work has shown manipulators to exploit online feedback using both numerical ratings and textual comments. The manipulation of ratings alone does not, however, affect book sales because buyers may discover these promotional steps. Nevertheless, handling in the form of a written part that represents a person's context, such as emotions, can affect the customers' decision considerably.

Scholars of [5] examined the identification of abnormalities based on the concept of separating

natural comparability from abnormal behavior. Throughout this study, an improved sentiment classification system for the identification of anomalies via social network analysis was developed successfully. Tweet data are a case study that demonstrates the effectiveness of the proposed method. According to implementation, the pattern of anomaly sentiment was successfully detected and interpreted. This work provides new ideas for the development of a comprehensive social media data system for examining feelings and trends of anomalies.

A study by [6] confirmed that modern applications use massive repositories to store their collected data every day. Nonetheless, only certain data are of interest and essential for evaluating and making important decisions based on information-rich data after processing. Regarding that, this paper proposed a data mining application that would help curb, consciously, and unknowingly, the impact of posting offensive articles or comments on the social network. The proposed method used in this study is based on the ID3 decision tree. This algorithm is performed with a custom algorithm that provides the table with a comment number and meaning for the ethnic, cultural, historical, and racist categories. This Table is then given to the algorithm of the decision tree ID3 that provides a decision to follow the derogations on social networking sites. The accuracy of the proposed algorithm is approximately 86%.

In [7] stated that the quantity and effect of online reviews continue to increase as the Web continues to expand in both scale and significance. Manufacturers and retailers are increasingly interested in consumer feedback and reviews due to their impact. An analysis of important machine learning techniques proposed to solve the spam detection problem is carried out.

In another study, [8], the question of identifying suspect blocks in multi-modal data is considered, although the data is incomplete. The authors proposed a method to measure block suspicion in a multi-modal tensor and demonstrated the satisfaction with axioms which delineate an effective metric as advanced by past research and also with the additional axiom. In this paper, they implemented CrossSpot+, which is an algorithm that was adapted from [9]. To uncover suspect groups in an incomplete multi-modal or partly paired results, experimental results suggest that CrossSpot+ manages this information efficiently and publishes signed results with a list deletion, which are substantially better than CrossSpot.

Authors of [10] carried out a thorough analysis using text analytics to identify suspicious users in social media in various existing systems. The techniques built in this area are based on text analysis so that a questionable content can be removed from suspicious activity regarding users' profiles on the site, and they also evaluate users' actions based on their user's details on the posts, comments, links exchanged and so on. Moreover, the writer mentioned the categorized group of suspicious behaviors like terrorist activity, financial laundering, and so on that would be possible to establish a corpus of potentially suspect terms that will further help build more detailed and accurate detection techniques. The outcomes of this study revealed that swarm-based search strategies would demonstrate their excellence and supremacy.

A study by [11] studied a comprehensive account of various existing methods for detecting suspicious user behaviors in online forums. The authors found that data mining is based on technologies built in this field to lessen the user's suspicious activity on the internet. This study shows that the user's behavior needs to be evaluated with input, suggestions, and input exchanged. Suspicious activities may also be categorized as terrorist, political, hacking, sexual and racial harassment classes, etc. This categorization can be used to create a corpus of potentially suspicious terms that help to establish more refined and trusted techniques to detect such activities.

In [12] stated that the conduct of disinformation users on social media sites is helpful in recognizing the templates and tactics used by misinformation agencies. In this paper, authors investigate the conduct of users on a popular social network platform, Reddit, designed as a forum for discussions on various subjects. Moreover, they present an initial and exploratory review of the actions and contents of 944 suspicious Reddit accounts, using the sentiment and activity analysis.

A study by [13] confirmed that the problem of bot detection had become an important field of study, as many businesses invest money in detecting suspicious behavior. As a result, multiple bot detection classifiers were suggested, but no proposal is made using understandable models for social network bots detection. For a social bot, specifically for Twitter, authors used a pattern- classification method. In addition, they are launching a new social bot identification feature platform that extends part of the current platform with Twitter-based usage functionality and tweet content analysis. The experimental results showed that the proposed methods provided better results compared with the results produced by other mechanisms published in the literature.

In [14], in their research, a sentiment-driven extremist classification system based on twitter posts users is implemented. Three modules of work are included: 1-tweeting, 2- preprocessing and 3- classification of extremist and non-extremist groups

using the LSTM+CNN model as well as other machine-learning and profound learning classifications. The experimental findings indicate that, in terms of greater accuracy, reminder, and f-measurement, the proposed algorithm surpassed the comparative methods. The algorithm, however, has certain limitations such as an Automated Crawling, Cleaning and Saving Method for Twitter content; 2- the lack of consideration of the visual and social context in order to achieve robust performance; and 3- the analysis of other extremisms using Deep Learning for multi-class mark classification Methods.

3. Methodology

Social media is a global networking platform for sharing a plethora of opinions, feelings, and emotions regardless of geographical, ethnic, and religious barriers. These views or sentiments can cover many hot topics such as education, politics, health, economics, etc. Accordingly, it becomes very critical to find out opinion sharing patterns and dig them to exploit the individuals' suspicious behaviors as a social media platform is a haven for these people to communicate with their companions. They use social media networks to address their extremist thoughts to other people as well. Thus, it is essential to analyze the opinions posted on the internet to classify them as good, bad, or neutral. Figure 2 illustrates the proposed methodology used to conduct this study. The full chain of the algorithm is developed using Python.
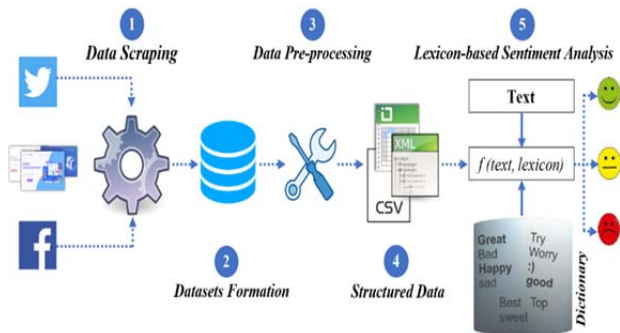

Figure 2. The proposed methodology of this study

3.1. Data Acquisition

3.1.1. Data Scraping

The data acquisition phase involves two basic steps that are data scraping and the formulation of a dataset. In the context of this study, the data was gathered from two networking sites that are Twitter and Facebook.

3.1.2. Formation of Datasets

To conduct this study, two datasets with English texts were used. The first dataset consists of 5,000 tweets collected from 50 random Twitter accounts. Whereas the second dataset comprises 12,000 posts gathered randomly from 50 Facebook public accounts.

3.1.3. Data Pre-processing

Data preprocessing is a pre-requisite step prior to data analysis as it filters the raw data (i.e., text) by eliminating symbols, irrelevant words, etc. to ensure data consistency. [13] states that data preprocessing "…is often the first step in the pipeline of a Natural Language Processing (NLP) system, with potential impact on its final performance." [13]. This step is of a great necessity, especially if the data required for analysis is a text taken from Facebook and/or Twitter, characterized as unstructured data. Generally, it aims at converting data into such a form that can be easily used for further processing [15]. In this study, Figure 3 [6] show an example of data preprocessing phase undergoes several steps involving: (a) text cleaning by eliminating noise such as HTML, hashtags (i.e., #BUiD), emoticons, punctuations along with other special characters, followed by (b) normalizing the text through removing blanks, stop words {to, all}, small words {It, was, my}, etc. Afterward, (c) text chunking, also known as shallow parsing, is applied using NLP based Part-Of-Speech (POS) tagging. Such a concept is utilized to classify different grammatical words into nouns, verbs or adjectives. At this stage, long strings text will be divided into different chunks which consist of different correlated phrases. Data chunking can also be effectively used for Name Entity Recognition (NER) to categorize texts into pre-defined classes, based on nouns phrases, particularly people, places, organizations etc. After the chunking process, a cleaned and filtered text is obtained and saved as a file in .CSV format for further sentimental analysis.
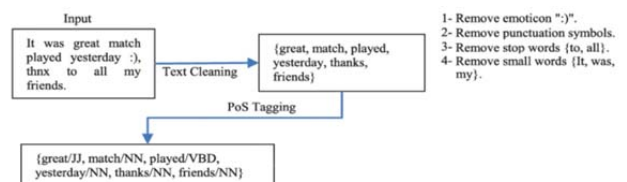

Figure 3. Data pre-processing steps [16].

3.2. *Sentiment Analysis*

Sentiment analysis also referred to as opinion mining, is a well-known approach for text analyses. It is defined by [16] as "…the interpretation and classification of emotions (positive, negative and neutral) within text data using text analysis techniques." Due to the dramatic growth of e-commerce and the huge demand on social media sites, people views and opinions with respect to the service, product or even thoughts of a particular topic can be easily identified through applying various techniques such as Natural Language Processing (NLP) and Machine Learning (ML). A wide range of studies has been conducted to analyze sentiments using the above-mentioned techniques. Generally, sentiment analysis can be carried out using the lexical and ML approaches, as shown in Figure 4. The lexical approach aims at mapping each word to sentiment through constructing a dictionary of sentiments without the need for pre-defined labels (i.e. samples). On the other hand, the machine learning approach requires training the model with pre-defined labels to classify the sentiment of newly input text.
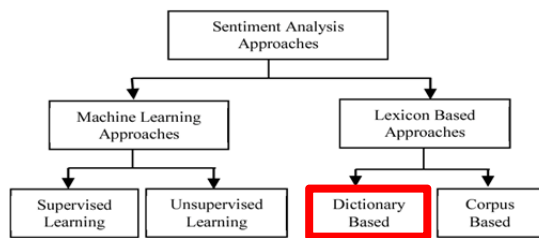

Figure 4. Sentiment analysis approaches

In this study, Valance Aware Dictionary Sentiment Reasoner (VADER), a lexicon-based sentimental analysis, is utilized. According to [16], VADER relies on "…a dictionary, which maps lexical features to emotion intensities called valence scores. The valence score of a text can be obtained by summing up the intensity of each word in the text." In the context of this study, VADER is applied to identify suspicious activities from texts posted on social media sites, particularly Facebook and Twitter, to classify sentiments into different categories of positive, negative, and neutral along with obtaining the intensity of emotions. The result of sentiments analysis can be classified into numerical data by finding the scores of an individual word or a complete sentence, as a whole, to carry out the classification (positive, negative, and neutral). The advantage of this approach is that there is no necessity of training the model with labeled data.

4. Results and Discussions

Social VADER sentiment analysis is represented by numeric scores, which are used to classify the results from the most negative to the most positive sentiments. Negative sentiment is an indication of having suspicious activity, whereas positive sentiment refers to the absence of suspicious activity. VADER purely depends on the lexicon feature dictionary, which helps to map the lexicon features such as words and sentences with sentiment intensities called valence scores. The overall score of a text is determined by adding up the intensities of all individual words within the text.

The major challenge in VADER is constructing a dictionary consisting of various features (i.e., lexicon features). Everything in a tweet and post is a lexicon feature which includes textual data as well as graphical objects smileys. VADER maps all these features with a dictionary, and a scale is used to report the results. This scale has values ranges from -4.00 to +4.00, where the negative four is called the most negative, and positive 4 indicates the most positive. The central point of the scale shows the impartial behavior of sentiments of the input text. Neutral behavior comes when the overall addition's sum is 0. Due to human cognitive behaviors towards a particular word, arbitrary judgments can be given. One word is more negative according to a certain person; nevertheless, the other's perception of the same word may be different, which affects the dictionary precision. If a dictionary is made based on false reviews, the results will be inaccurate. To overcome this problem, VADER sentimental analysis relies on multiple reviewers, calculating the average ratings for each word.

The VADER uses a different scale for the sentence based features mapping, from -1 (most negative) to +1 (most positive) that is different from the ordinary word-based feature mapping. According to [16], the score of a sentence is obtained by "… summing up the valence score of each sentiment-bearing word". Finally, normalization is applied to get the mapped result within the interval of [-1, +1]. This normalization is called "Hutto Normalization," which can be calculated as per equation (1).

$$N = \frac{x}{\sqrt{x^2 + \alpha}}$$

(1)

Where N is a normalization value, x is the sum of sentiment score of all words of a sentence, and alpha (α) is a constant, called normalization parameter. In this study, α = 15, as suggested by [17]. Based on the literature, VADER proved to be a beneficial approach in the analysis of short texts such as tweets, sentences, comments, and small posts.

Figure 5. Sample results (Twitter data).

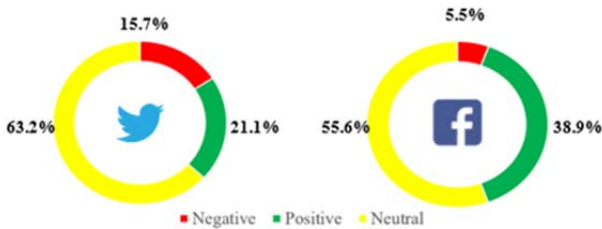Figure 6. Sample results (Facebook data)

Figure 7. Sentimental scores (positive, negative, and neutral responses) based on Twitter and Facebook posts
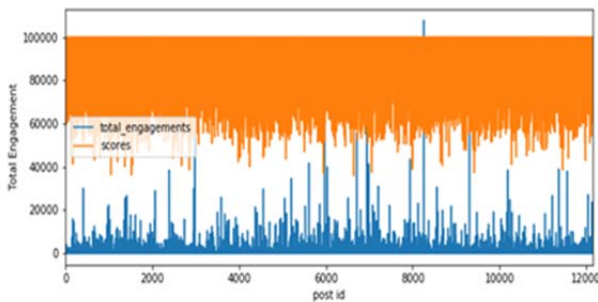

Figure 8. Total engagement of every post along with the scores of sentiment (Facebook)
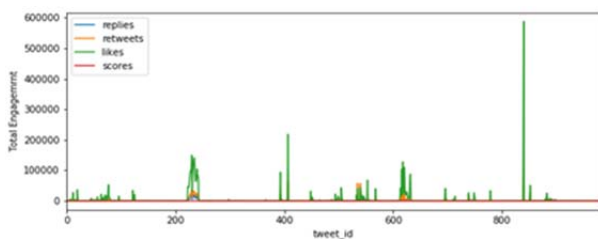

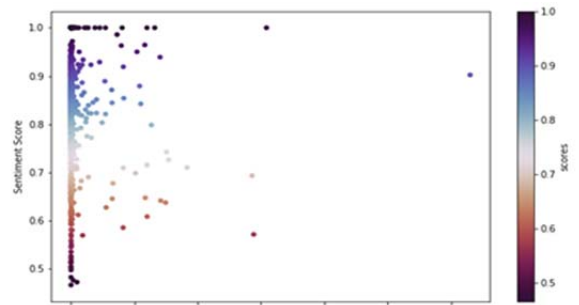Figure 9. Total engagement of every post along with the scores of sentiment (Twitter)

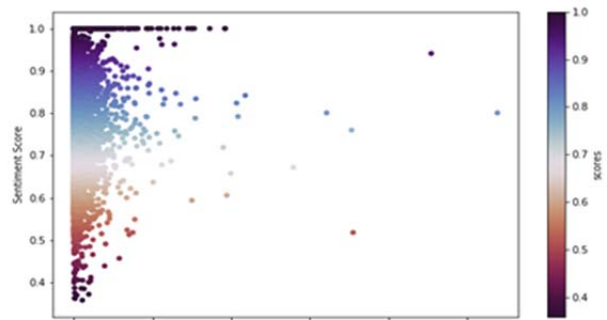Figure 10. Intensities of sentiments (Twitter data)

Figure 11. Intensities of sentiments (Facebook data)

5. Conclusion

Due to the advancement of mobile technologies and the universal accessibility of the internet, a large chunk of prejudiced content has been posted on the social networking platforms such as Twitter and Facebook, on daily basis. Accordingly, social media sites are full of opinions, thoughts and emotions conveying many messages that can be suspicious. This requires a lot of attention and frequent monitoring because these sentiments may lead to unexpected serious actions. In this study, data was collected from two platforms, Twitter and Facebook, resulting in the creation of two datasets. The first dataset consists of 5,000 tweets while the second one comprises 12,000 posts obtained from the top public Facebook accounts. Prior to the implementation of sentiment analysis, preprocessing steps are applied to ensure data consistency by eliminating noise (HTMLs, hashtags, emoticons, punctuations), normalizing the text as well as chunking the text using NLP based Part-Of-Speech (POS) tagging. Afterwards, Valance Aware Dictionary Sentiment Reasoner (VADER), a lexicon based sentimental analysis, is utilized to identify the suspicious activities from texts posted on social media sites particularly Facebook and Twitter; to classify sentiments into different categories of positive, negative and neutral along with obtaining the intensity of emotions. The advantage of this approach is that, there is no necessity of training the model with a pre-defined labeled data, which is the case in machine learning. By analyzing the initial results, it

has been noticed that VADER works very well for the smaller size texts such as comments and tweets, etc. The following findings have been pointed out from this study: Initially, VADER twitter sentimental analysis showed that out of 5000 tweets 50.8% people shared a neutral opinion, 39.2% shared negative opinion and only 9.9% showed positive opinion. Secondly, on Facebook, the majority of people showed a neutral response which is 55.6%, 38.9% shared positive response and only 5.5% shared negative opinion. Thirdly, the score of sentiments and engagement in every post affects the intensities of sentiments. The proposed model can be further enhanced by collecting more and more data, for training, to increase the efficiency of the system, and then it can predict the suspicious activity.

References

[1]. Rodermund, S. C., Lorig, F., Berndt, J. O., & Timm, I. J. (2017, August). An agent architecture for simulating communication dynamics in social media. In *German Conference on Multiagent System Technologies* (pp. 19-37). Springer, Cham. https://doi.org/10.1007/978-3-319-64798-2_2

[2]. Kim, D. (2016). Value ecosystem models for social media services. *Technological Forecasting and Social Change*, *107*, 13-27. https://doi.org/10.1016/j.techfore.2016.03.010

[3]. Santos, V., & Gonçalves, R. (2012, September). A Social Shopping Model Proposal. In *MCIS (Short Papers)* (p. 20).

[4]. Hu, N., Bose, I., Koh, N. S., & Liu, L. (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision support systems*, *52*(3), 674-684. https://doi.org/10.1016/j.dss.2011.11.002

[5]. Wang, Z., Joo, V., Tong, C., Xin, X., & Chin, H. C. (2014, December). Anomaly detection through enhanced sentiment analysis on social media data. In *2014 IEEE 6th international conference on cloud computing technology and science* (pp. 917-922). IEEE. DOI: 10.1109/CloudCom.2014.69

[6]. Pandhe, S., & Pawar, S. (2015). Algorithm to Monitor Suspicious Activity on Social Networking Sites using Data Mining Techniques. *International Journal of Computer Applications*, *116*(12).

[7]. Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., & Al Najada, H. (2015). Survey of review spam detection using machine learning techniques. *Journal of Big Data*, *2*(1), 23.

[8]. Chiu, C., Zhan, J., & Zhan, F. (2017). Uncovering suspicious activity from partially paired and incomplete multimodal data. *IEEE Access*, *5*, 13689-13698. DOI: 10.1109/ACCESS.2017.2726078

[9]. Hailong, Z., Wenyan, G., & Bo, J. (2014, September). Machine learning and lexicon based methods for sentiment classification: A survey. In *2014 11th web information system and application conference* (pp. 262-265). IEEE. DOI: 10.1109/WISA.2014.55

[10]. Asha, A. & Balkishan, D. (2017). Systematic Review on the Suspicious Profiles Detection on Online Social Media Data. *Oriental journal of computer science and technology, 10*(3), 644-652.

[11]. Srivastava, T., Mangalagowri, R., & Dudala, S. S. (2018). Monitoring of Suspicious Discussions on Online Forums Using Data Mining. *International Journal of Pure and Applied Mathematics*, *118*(22), 257-262.

[12]. Eunjin Lee et al. (2019). An Exploratory Analysis of Suspicious Reddit User Accounts based on Sentiment and Interactions. Retrieved from: https://dais-ita.org/sites/default/files/3926_paper.pdf [accessed: 10 May 2020].

[13]. Camacho-Collados, J. and Pilehvar, M.T. (2018). On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics. pp. 40–46

[14]. Ahmad, S., Asghar, M. Z., Alotaibi, F. M., & Awan, I. (2019). Detection and classification of social media-based extremist affiliations using sentiment analysis techniques. *Human-centric Computing and Information Sciences*, *9*(1), 24.

[15]. Fouad, M. M., Gharib, T. F., & Mashat, A. S. (2018, February). Efficient twitter sentiment analysis system with feature selection and classifier ensemble. In *International Conference on Advanced Machine Learning Technologies and Applications* (pp. 516-527). Springer, Cham. https://doi.org/10.1007/978-3-319-74690-6_51

[16]. Borg, A., & Boldt, M. (2020). Using VADER sentiment and SVM for predicting customer response sentiment. *Expert Systems with Applications*, *162*, 113746.

[17]. Gilbert, C. H. E., & Hutto, E. (2014, June). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*.