International Research Journal of Education and Technology
ISSN 2581-7795

Data Mining to Track Suspicious Discussions on Online Forums

Sheikh Md Zubair Md Zahoor

Former Research Scholar, Computer Science, OPJS University, Churu, Rajasthan, India

---

Abstract - *Since we all aware, people use the internet to access a variety of social media platforms and other platforms for various purposes. Over networks, massive amounts of data are transmitted. The internet has made online communication and business exceedingly simple and quick. People use the internet for a variety of reasons all across the world. Whereas the internet is utilized for positive purposes, it is sometimes used for harmful or unlawful objectives. These sites are also utilized for a variety of criminal acts such as terrorism, threads, copyright violations, phishing scams, frauds, and spams, among others. Various strategies are being used by law enforcement agencies and departments to address these issues.*

Key Words: Stop word • Stemming algorithm • Suffix & Affix Stemmers • Levenshtein algorithm • Classification • Brute Force Algorithms • SNS • OSMS

1. INTRODUCTION

Fast growing IT and communication technologies are providing lot of different online forums for communication. There are also many forums that are used unlawfully for illegal activities that are really threat for the society. Lot of malicious people are usingthese forums for criminal purposes. People discuss about many different entities on online forum in different forms (text, video, photos etc.) without any restriction. Peoplegive feedback and rank entities and exchange information openly without any rule andregulation. By distributing textual material through a browser interface, the web has become a very convenient and effective communication medium for individuals to share their knowledge, express their opinions, advertise their products, and even teach each other. It is critical for humans to extract relevant information from plain textual material in order to discover the hidden facts. Data mining's major goal is to extract information from massive data sets and put it into a format that can be understood. As Internet technology has advanced, it has spawned a plethora of both legal and illicit operations. It has been discovered that much first-hand news is debated on Internet forums long before it is publicized in traditional media. This communication channel serves as an efficient conduit for unlawful operations such as the distribution of copyrighted films, threatening messages, and online gambling, among others. Rapid advancements in information and communication technology have ushered in new avenues for online debate while simultaneously narrowing the gap between individuals. Unfortunately, malevolent people take advantage of this technology for unlawful objectives. People employ a variety of suspicious letter types (text, image, video, etc.) and many online output formats to trade suspicious messages on social plate forums. Because most social networking sites use textual data, we'll just focus on the text letters. The text mining technique is a good way to bring semantics to a key part of the study problem. Text analysis is used in the same way to discover questionable social media posts. Data mining techniques will be used to extract meaningful information from user-posted content on the internet. The internet provides a global forum for the exchange of ideas. The number of users who provide feedback on items is continually increasing. For market research, online opinion analysis is an important technique. The importance of an automatic mode of thinking is frequently emphasized. In this research, we use a Text mining approach to collect, aggregate, and track customer comments. This technique to study and commercial application compatibility is demonstrated in the sports business. Web-crime is growing more ubiquitous as high-speed Internet becomes more popular, and the majority of it is in textual form, because the majority of illicit or illegal information in documents is defined by occurrences. Web-based crime patterns and trends can be identified using Event Based Linguistic Technology. This work attempts to provide a review of data mining for the extraction of useful information.

Terrorism and illegal activities have been dealt with for a long time through social networking sites (SNS) and online social media sites (OSMS) because there are numerous social networking sites (SNS) such as Google plus, Twitter, Facebook, and Yahoo that provide a platform for everyone to communicate and share information over the internet. As a result, a large number of these platforms lack strict standards and laws for transmitting information over the network. As a result, these platforms can readily be used for illicit or terrorist actions. People can simply create groups and communities to discuss any type of information, as well as start any type of complaint, using these platforms. OSMS and SNS are cited in a study by the US Army as being crucial in the implementation of terrorist actions. SNS and OSMS are used to carry out over 90% of terrorist acts. Law enforcement agencies now employ OSMS and SNS to conduct investigations. We are employing NLP and LSA systems to detect these suspicious actions.

2. LITERATURE REVIEW

Different strategies were utilized in the research Automated Monitoring Suspicious Discussions on Online Forums Using Data Mining Statistical Corpus Based Approach to identify suspicious talk on online forums. These are the following:

Stop Words Selection

Stop words are regularly used terms in the English language, such as the pronouns "he, she, it, they, we" or "the, an, a" or prepositions articles. The Information Retrieval System was the first to use these terms. In terms of frequency of small size, beautiful words in the English language made for a significant portion of the text. Because these words are frequently used in the English language, it was assumed that they do not convey useful information. The application-independent stop words list has been eliminated. This text mining could be unfavorable to mining applications.

Brue Force Algorithms

The stemmers table shows how root forms and inflected forms are related. To avoid a term, the table is requested to find a matching inflection. If a match is detected, the root form associated with it is returned.

Stemming Algorithm

Although the stemming algorithm does not rely on a lookup table, it does contain some principles that are used to determine the root form of a particular input word. There are some guidelines to follow.

- Remove the ness from a word if it is happiness.
- Remove ful if it's thankful.
- If it's serving then delete the ing, etc.

Stemmers with Suffixes and Affixes

The word affix can refer to either a prefix or a suffix in linguistic terms. Removing prefixes from suffixes is handled using a variety of approaches. When a prefix "dis" is used with a phrase like "disappear," or when a prefix "en" is used with a word like "enclose," these prefixes can be removed. By applying the same methods as mentioned earlier is called as many, and affix stripping. This formula (MWC = BS/AS) is used to calculate Stemmer Strength, while this formula (ICF = (BS AS)/BS) is used to calculate index compression. [10] The flow of the suggested framework is depicted in Figure 1.

Emotional Algorithms

Emotional algorithms are used to determine people's feelings and opinions on entities on online platforms using various types of data such as text, audio, and video provided by different users regarding the entity. Typically, these algorithms are used to detect emotions in text. The approaches listed below are used to identify emotions in text.

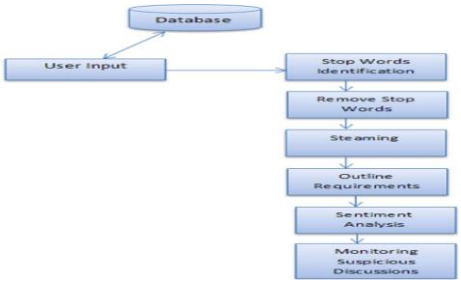

Figure 1: Proposed framework

Keyword Spotting Technique

The nature of emotion is defined as the output, and text data is used as the input. The first stage in this procedure is to turn text data into tokens, which are then utilized to detect emotion words. The type of feeling is identified after examining and evaluating the token words, whether it is positive, negative, or normal.

Learning-Based Methods

The way you learn base technique is a little different. We determined emotions from the input data in the previous stage, however in this procedure, emotions are classified according to their nature.

Hybrid Methods

Keyword-based and earning-based strategies that were previously used cannot produce totally satisfactory outcomes. To improve the outcomes, several systems combine the two processes to produce more accurate and satisfying results.

Text data mining techniques are utilized to mine the given data in the next work, Surveillance of Suspicious Discussions on Online Forums. These are the following:

Stop Word Selection

Stop words are the most commonly used words in the English language, and they include pronouns like "I," "he," and "she," as well as articles like "a," "an," and "the" and prepositions. The concept of stop-words was first established in the Information Retrieval (IR) system. These terms are eliminated from the submitted text data since they do not produce useful information.

Stemming Algorithm

The root word is derived from the input word during the stemming process. The root word "Jump," for example, is formed from the string "jumped".

Levenshtein Algorithm

It's used to figure out how far apart two words are. If the words are dissimilar, the distance will be high, but if they are similar, it will be zero or possibly the smallest value. We used the Twitter dataset and assessed it using these techniques for spotting suspicious profiles on social media sites.

Text Corpus

On many social media sites, a large volume of material is uploaded. We can acquire data using a variety of methods. We're going to use the Twitter dataset for this. In May of

2011, it was collected. There are three million user profiles, fifty million tweets, and 284 million followers on Twitter.

Corpus Processing

Data is pre-processed in this step. The root words are extracted via stemming, and the stop word technique is utilized to remove extraneous terms from the input data.

Classification

Two words are matched at this stage, and the results are then classified. Words that are similar or similar in meaning are stored in the same group, whereas separate words are stored in a different class. It's difficult to compute text form data to determine similarity. Manhatton, Minkawski, and Euclideant distance formulae are used for numeric data [12], and Jaccuard and Haming distance formulas are utilized for groups or communities [17–19].

The proposed system is depicted in Fig. 2 below.

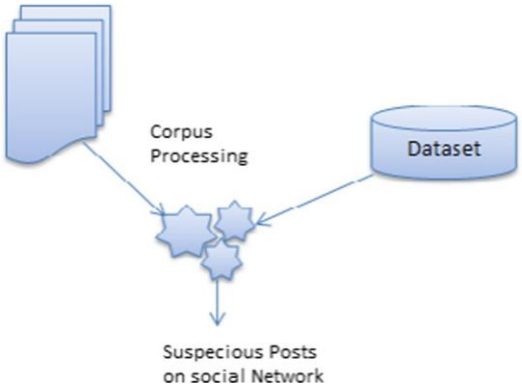

Figure 2: Proposed System

3. MATHEMATICAL FORMULATION

The similarity between the words used in the tweets (a) and the suspicious word (b) in the database was determined using Normalized Compression Distance.

NSDðx; yÞ ¼ CðxyÞ — minfCðxÞ; CðyÞg=maxfCðxÞ; CðyÞg (1)

Where

$$0 \le NCD(x, y) \le 1 \quad (2)$$

First a post is decomposed then NCD is executed. If NCD (a, b) = 0, then word “a” and “b” are same and the words are different if NCD (a, b) = 1. Objects are classified on the bases of distance.

4. EVALUATION

Strings that are discovered to be similar are stored in the suspect class.dfv file (Tables 1 and 2).

Term 1	Term 2	NCD
Explosion	Explosion	0
Terrorist	Terrorist	0
Attack	Attack	0

Table 1: Results of NCD calculating between similar words

Term 1	Term 2	NCD
Make	Explosion	0
Internet	Terrorist	0
Use	Attack	0

Table 2:Results of NCD calculating between different words

Clustering classifies things based on shared characteristics. Objects with similar attributes are grouped together in a class, whereas objects with distinct properties are divided into two groups.

Then, predefined objects are linked to classified objects. Link analysis is used to identify linkages between things and to derive frequently occurring objects. In order to detect intruders in the network, the patterns formed by these items are also studied. Objects that occur at different times on a consistent time interval are crucial to examine. For reliable findings, we'll need carefully structured data.

After the outliers have been removed, classification is carried out. Objects or entities with similar qualities are gathered and organized into predefined classes in this process. The relevance of collected patterns is now reflected in the object's class.

For a long time, SNS and Online Social Media Sites (OSMS) have been used to combat terrorism and illicit acts. [25] In this research, we present a system that can monitor any information passing across a network between users. It can identify persons or groups of users in a group or social network who are behaving in an unusual or inconsistent manner. We all know how difficult it is to obtain data from any social media platform. It is only available to law enforcement agencies and authorities. For this project, we created "Manipal Net," a private social network. The five steps in our suggested method are also represented in Figure 3.

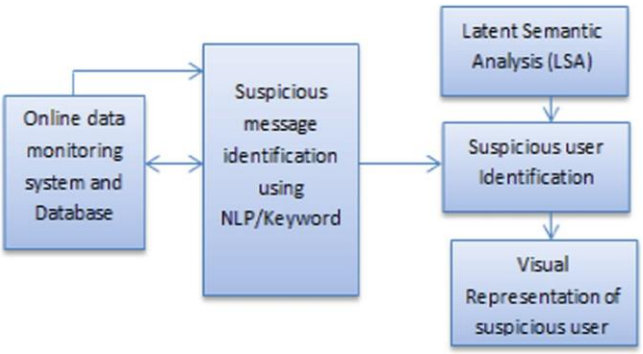

Figure 3: Using an NLP/keyword system, identify suspicious messages.

Suspicious messages are detected in this module of the proposed system. We found the suspicious messages by matching them to predefined items such as Hate Messages, Terrorist Activity, Delhi Gang Rape and Harm to Society, 'Narendra Modi,' and other Confidential Keywords. The following sub-modules make up this module:

Figure 4 depicts the use of NLP to identify suspicious messages.

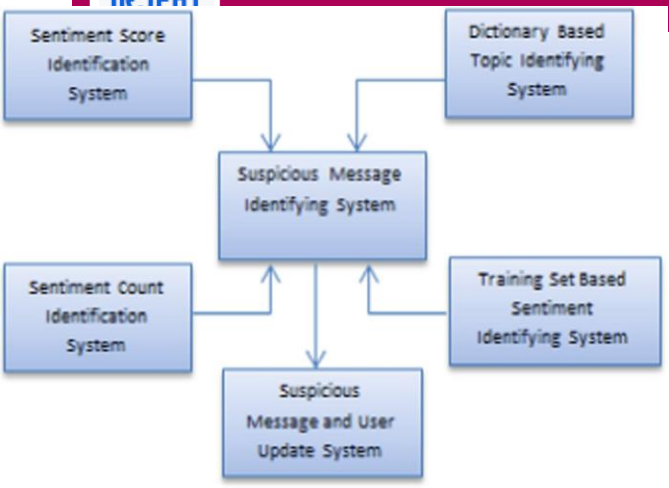

Figure 4: NLP-based detection of suspicious messages

Sentiment Score Identification System: This sub-module calculates a sentiment score for each message word. To score the words, a dataset from SentiStrength was used. The dataset contains 2546 words, and each word is given a score ranging from 1 to 5. Positive words receive a score of +1 to +5, while negative words receive a score of 1 to 5.

Sentiment Count Identification System: Based on a set of 3905 negative and 2230 positive predefined words, this module determines the likelihood of negative and positive sentiment words in a message.

Sentiment Identifying System Based on a Training Set: In this module, the sentiment of the message is determined using a method known as "sentiment analysis" on the basis of a training dataset. Training datasets are gathered from a variety of sources [30–33]. It makes use of the "Sentiment Score Identification System" module's output.

Topic Identifying System Based on a Dictionary: In this module, a topic dictionary is employed to assign a unique matching score to each topic.

Following the execution of the processes, a suspicious message and its corresponding user were identified. We classified the communication as either normal or suspicious using NLP [34, 35].

Latent Semantic Analysis (LSA) System

This module of the system uses Latent Semantic Indexing (LSI) and Singular Value Decomposition (SVD) to identify groups of users. Messages that are similar across networks are aggregated and examined. If a suspicious message is discovered, and another communication in the network is discovered to be similar to the suspicious message, even if their words are not identical, these messages are grouped together as they are discussing a single common issue. Using these aggregated communications, a suspect user group can be quickly identified.

Suspicious Users' Identification System

The odds of making a mistake are reduced in this module. A suspected user's history is kept, and after some time, current information and the user's past history are examined, and a final conclusion is reached in a sub section called "suspicious user alert." Figure 5 depicts a visual representation of users who are suspicious.

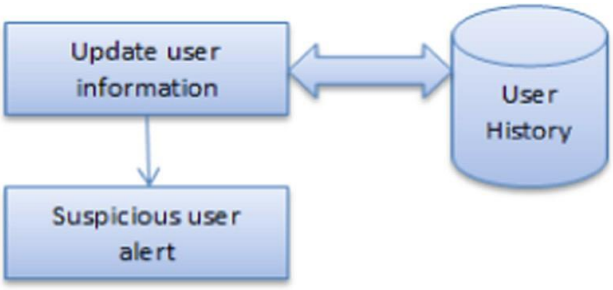

Figure 5: Visual representation of suspicious users.

After identifying suspicious individuals on the network, these users are highlighted on the network using Gephi, a visualization tool. Gephi graphically depicts the network, which is made up of nodes and edges. It also draws attention to a group of people who are discussing the same subject. Figure 6 shows Manipal network diagram.

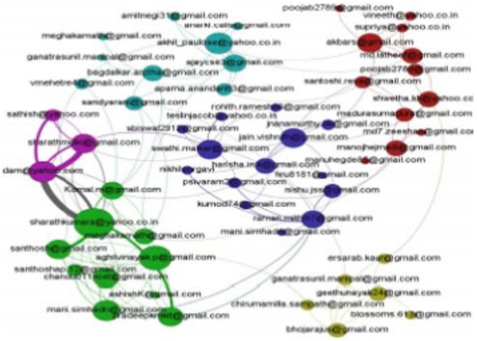

Figure 6: "Manipal network" Diagram

5. DISCUSSION

The result displays only suspicious words. It does not specify the context in which the terms are employed. We must adhere to a predefined suspicious word list in the suggested system. The proposed approach can only inform us which words are suspect; it cannot tell us if they are utilized favorably or negatively. In the proposed solution, if a term seems suspicious but isn't on the predefined list, the algorithm will overlook it, which might represent a major hazard. The suggested method employs previously employed mining techniques, but it approaches the dataset in a unique way. It finds the distance between the root words and the predefined suspect terms and classifies the words using mathematical algorithms. However, it also declares the terms suspicious when they are combined with suspicious words, which is ineffective. To begin with, practically every paper ignores the context in which the phrases are used. This, I believe, is the most effective method for detecting suspicious terms in online forums. It also identifies

suspicious people based on the user's present and prior use of suspicious terms. It's also capable of detecting suspicious groups. People who discuss common suspect themes and use suspicious terms are gathered and analyzed in groups. The most effective aspect of the suggested system is that it does not flag a user as suspicious; instead, it builds up a history of the user and analyzes that past as well as the phrases he or she uses. A person is classified as suspicious or not suspicious based on previous data, and people connected to the user are identified and a group is classified as suspicious.

6. CONCLUSIONS

Root words are taken from the data and extraneous material is ignored in this article, which takes a concentrated approach. Processing time is cut in half, and costs are cut as well. Different emotional approaches are also used to identify user feelings. Another strategy is to increase the frequency of terms in order to improve the effectiveness of the outcomes. However, it does not specify how these ominous terms are employed. For what reason has the user used the phrases in a good or negative sense? In the proposed solution, a focused approach is adopted, and irrelevant material is excluded. A set of words is matched with a set of predefined words. However, there is no context in which the word is missing. If a suspicious term does not appear in the predefined list, it is ignored. Root words are retrieved and superfluous words are avoided in this strategy, which saves time and effort. The suggested method is utilized with the Twitter dataset, and words are classified by matching the pre-defined list. Not only are suspicious terms classified, but words that are related to them are also classified as suspicious. It has presented many solutions for various types of data, such as numeric, text, and group data. However, we must define the suspicious terms list according to our needs at all times. The words' context is ignored. Words which are used in conjunction with suspect words are also labeled as suspicious, which is not a scientific technique. The user id and suspicious terms used by users are kept and analyzed in a database. A user is not labeled as suspect because of a single dubious term, but rather because of a succession of words used by that user. However, the context in which the questionable words are employed is also ignored in this case. And there is a pre-defined list of suspicious words that is not general but specific. As a result, for each new scenario, a separate list is required based on the situation and requirements.

REFERENCES

[1] Murugesan, M.S., Devi, R.P., Deepthi, S., Lavanya, V.S., Princy, A.: Automated monitoring suspicious discussions on online forums using data mining statistical corpus based approach. Imp. J. Interdiscip. Res. 2(5) (2016)

[2] Upganlawar, H., Sambhe, N.: Surveillance of suspicious discussions on online forums using text data mining. Int. J. Adv. Electron. Comput. Sci. 4(4) (2017)

[3] Alami, S., Beqqali, O.E.: Detecting suspicious profiles using text analysis within social media. J. Theor. Appl. Inf. Technol. 73(3) (2015)

[4] Kaiser, C., Bodendorf, F.: Monitoring opinions in online forums-a case study from the sports industry. Int. J. Inf. Educ. Technol. 2(3), 212 (2012)

[5] Hosseinkhani, J., Koochakzaei, M., Keikhaee, S., Naniz, J.H.: Detecting suspicion information on the Web using crime data mining techniques. Int. J. Adv. Comput. Sci. Inf. Technol. 3(1), 32–41 (2014)

[6] Yao, Z., Ze-wen, C.: Research on the construction and filter method of stop-word list in text preprocessing. In: Proceedings of 2011 IEEE Intelligent Computation Technology and Automation (ICICTA), pp. 217–221, 11–13 (2011)

[7] Ayral, H., Yavuz, S.: An automated domain specific stop word generation method for natural language text classification. In: International Symposium on Proceedings of Innovations in Intelligent Systems and Applications (INISTA), pp. 500–503, 15–18 June 2011

[8] Silva, C., Ribeiro, B.: The importance of stop word removal on recall values in text categorization. In: 2003 Proceedings of the International Joint Conference on Neural Networks, vol. 3. IEEE (2003)

[9] Yu, S.: Stemming algorithm for text data and application to data mining. In: Proceedings of 2010 IEEE 5th International Conference on Computer Science & Education (ICCSE), pp. 507–510, 24–27 (2010)

[10] Porter, M.F.: An algorithm for suffix stripping. Program 14(3), 130–137 (1980)

[11] O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From tweets to polls: linking text sentiment to public opinion time series. In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (2010)

[12] Ho, T.K.: Stop word location and identification for adaptive text recognition. In: Proceedings of 2000 IEEE International Journal on Document Analysis and Recognition, vol. 3, no. 1 (2000)

[13] Zeng, Z., Yang, H., Feng, T.: Data mining methods for knowledge discovery. In: Proceedings of 2011 IEEE International Conference on Data Mining Methods for Extraction of Data, pp. 412–415, 29–31 (2011)

[14] Yang, Y.: An evaluation of statistical approaches to text categorization. In: Proceedings of 1999 IEEE Journal on Information Retrieval, vol. 1, no. 1 (1999)

[15] Li, R., Wang, S., Deng, H., Wang, R., Chang, K.C.-C.: Towards social user profiling: unified and discriminative influence model for inferring home locations. In: KDD 2012, Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA (2012)

[16] Porter, M.F.: An algorithm for suffix stripping. Program 14(3), 130–137 (1980)

[17] Marquiz, S.: Classificateur de Kolmogorov sur le web 7 Juin (2004)

[18] Levorato, V., Van Le, T., Lamure, M., Bui, M.: Distance de compression et classification prétopologique (2009)

[19] Kaufman L., Rousseeuw P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley Interscience (1990)

[20] Dommers, M.: Calculating the normalized compression distance between two strings, 20 January 2009

[21] Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Surv. 31(3), 264–323 (1999)

[22] Agrawal, R., Imielinski, T., Swami, A.N.: Mining association rules between sets of items in large databases. In: Proceedings of the ACM SIGMOD International Conference on Management of Data (1993)

[23] Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco (2009)

[24] Agrawal, R., Srikant, R.: Mining sequential motifs. In: 11th International Conference on Data Engineering (1995)

[25] Frank, R., Cheng, C., Pun, V.: Social media sites: new fora for criminal, communication, and investigation opportunities. Research and National Coordination Organized Crime Division Law Enforcement and Policy Branch Public Safety Canada (2011)

[26] Alderson, M.: Facebook: a useful tool for police? Connectedcops. 25 January 2011. Web, 3 February 2011

[27] Sentistrength - sentiment strength detection in short texts. http://sentistrength.wlv.ac.uk

[28] Caren, N.: An Introduction to Text Analysis with Python. http://nealcaren.web.unc.edu/

[29] Gokulakrishnan, B., Priyanthan, P., Ragavan, T., Prasath, N., Perera, A.: Opinion mining and sentiment analysis on a Twitter data stream. In: 2012 International Conference on Advances in ICT for Emerging Regions (ICTer), pp. 182–188 (2012)

[30] Recorded future: Creating an insightful world. https://www.recordedfuture.com/

[31] Voices of the Mumbai terror siege: Police taped chilling phone conversations between suicide terrorists and their Pakistani handlers. http://transcripts.cnn.com/TRANSCRIPTS/0911/15/fzgps.01.html

[32] The Hindu: Audio of 26/11 tape: Zabiuddin ansari briefs terrorists. http://www.thehindu.com/news/resources/article3568903.ecel

[33] Black Friday: The shocking truth behind the 1993 Bombay blast film conversation subtitle. http://www.subtitles.net/en/ppodnapisi/podnapis/i/206775/black-friday-2004-subtitlesl

[34] Jurafsky, D., Bethard, S.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Pearson Educa- tion Inc. (2009)

[35] Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. 1005 Gravenstein Highway North. O Reilly Media, Inc. Sebastopol (2009)

[36] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. 41(6), 391–407 (1990)

[37] Manning, C.D., Raghavan, P., Schutze, H.: Introduction to Information Retrieval. Cambridge University Press, New York (2008)

[38] Gephi: Network analysis and visualization. https://gephi.org/

[39] Kumar, A.S., Singh, S.: Detection of user cluster with suspicious activity in online social networking sites. In: 2013 2nd International Conference on Advanced Computing, Networking and Security (ADCONS), pp. 220–225. IEEE (2013)

[40] Bavane, A.B., Ambilwade Priyanka, V., Bachhav Mourvika, D., Dafal Sumit, N., Fulari Priyanka, Y.: Monitoring suspicious discussions on online forum by data mining