Journal of Discrete Mathematical Sciences & Cryptography
ISSN 0972-0529 (Print), ISSN 2169-0065 (Online)
DOI : 10.1080/09720529.2022.2072421

Taylor & Francis
Taylor & Francis Group

COVID-19 pandemic & cyber security issues : Sentiment analysis and topic modeling approach

Sonal Khandelwal [†]
Aanyaa Chaudhary *
TAPMI School of Business
Manipal University Jaipur
Jaipur-Ajmer Express Highway
Jaipur 303007
Rajasthan
India

Abstract

The purpose of this study is to understand public awareness of cybersecurity-related issues and discussions during the COVID-19 pandemic. Employees and confidential organizational data are vulnerable to cyber-attacks as a result of the pandemic, which has raised concerns about cyber security about the new normal of working from home. The public's main sentiments and aspects related to cyber security concerns have been mined from tweets on the microblogging social media website Twitter. Sentiment Analysis and Topic Modeling techniques have been applied to understand the perspectives, emotions, and themes discussed by people. The analysis reveals people are becoming more aware of cybercrime-related concerns and are more positive than negative in combating the challenge of cybercrime. The paper also highlights the main themes revealed by Topic Modeling.

Subject Classification: *68T35.*

Keywords: *Cyber security, Cyber crime, Sentiment analysis, Topic modeling, Text mining, Big data, Twitter.*

1. Introduction

The Internet has now integrated into human lives, with about 3 billion users all over the world. Billions of dollars are generated globally by this global network [13]. In the current scenario, the majority of the

[†] *E-mail:* sonal.khandelwal.sharma@gmail.com
* *E-mail:* aanyaa1912@gmail.com (Corresponding Author)

activities and interactions, whether commercial, cultural, governmental, nongovernmental, individual, or organizational, are all performed in cyberspace [1], causing unprecedented growth in cyberspace [2]. The innovations in the field of the internet have made it more accessible, user-friendly, and affordable for most people [34], thus integrating it into human lives profusely. The lives of humans on earth are intertwined in cyberspace and any turbulence or insecurity impacts different facets of human lives [18]. Such a tremendous growth in information sharing provides a platform to those with malevolent intentions and raises cyber-security issues [15]. Due to COVID-19 pandemic role of technology has escalated in the professional lives with remote working becoming new normal. However organizations are still not able to provide cybercrime safe remote working. In today's contemporary world social media is a great source of exchanging information, opinions, and discussions by providing a universal platform. Twitter, a microblogging platform has 206 million daily active users enabling users to create and post messages and follow each other [4]. People post their experiences, opinions especially emotions via electronic media. Therefore, microblogging websites like Twitter are a momentous source of data followed by opinion mining and sentiment analysis [17, 29, 30].

The study examines tweets related to cyber security. The paper tries to capture the tone of people being exposed to cybercrime and its impact on the professional and personal life of people during COVID 19.

2. Related Works

Cyber-Security

The majority of every country's financial, socio-cultural, politico-legal, and economic activities are now conducted in cyberspace at all levels [19]. Cyber attacks aim to cause economic impairment to businesses [19]. Cyber security denotes the maintenance of the Integrity, Confidentiality, and Availability (ICA) of computing possessions of an organization or organization's network with other organizations [15]. The work-from-home (WFH) business model raises cyber-attack and also poses threat to corporate data which is displaced to an online environment [26].

Social Network Analysis and Twitter as a source

Microblogging is an important platform generating surplus posts every day and sharing opinions from internet users [23]. Collecting and

analyzing data from researchers in social networks collects important information by examining different aspects of user behavior [21]. Social Network Analysis (SNA) is a scientific method for quantitatively and qualitatively extracting and analyzing data and its structural features. The data generated from Twitter is regarded as very crucial and beneficial for research accepted universally [6]. People have expressed their feelings on natural disasters previously including SARS and COVID -19 is not an exception [22, 28].

Topic Modeling Using Big Data Latent Dirichlet allocation (LDA) and Sentiment Analysis

The topic modeling analysis involves a set of techniques for identifying what the text is [3]. Topic modeling is a technique for extracting potentially important topics from a large number of documents and proposing them based on a procedural probability distribution model [14]. LDA approach has been used on online reviews and transportation research also. [16, 33]

In recent years there is a plethora of text data where people have expressed their opinions, thus text contains subjective elements profusely [9]. As a result sentiment analysis or opinion mining is gaining more and more significance to extract users' sensibility and utilize it [11]. Sentiment Analysis incorporates a series of systematic techniques where emotions are identified, classified, and quantified[27,3]. Sentiment analysis incorporates a sentiment dictionary that comprises emotional words along with their polarities. The polarity indicated the degree of positive, negative, or neutral words. Opinion Lexicon was built by [20] that comprise 4783 negative words and 2006 positive words in the English language. Sentiment categorization can be done either by machine learning or lexicon-based technique [7].

3. Methodology

Data Source

The data for the study is collected from Twitter and contained 24092 tweets from 26 November 2021 to 12 December 2021. Twitter application programming interface (API) keys were used to collect cybersecurity-related tweets in the English language.

The unstructured data was converted into a document term matrix, thus creating a structured format for analysis, and pre-processing was done subsequently. The detailed procedure adopted in the study is depicted in Figure1.

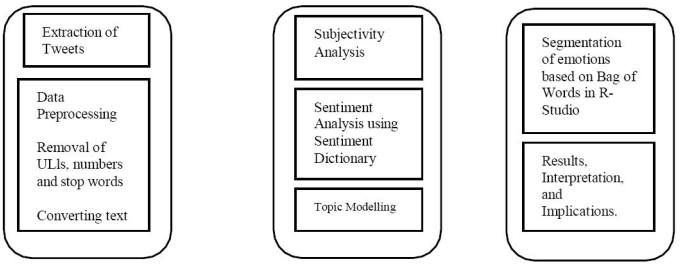

Figure 1

Overview of Research Flow

Preprocessing of Data

NLP tools and packages were used for cleaning data in Python and R. The process involved converting the text to lower case, removal of hyperlinks and URLs, removal of punctuations, and stop words. This involved stemming and lemmatization methods. The white spaces created are also removed to enhance classification performance. The stemming and lemmatization target reduction of inflectional forms and word's derivationally related forms are converted to common base forms occasionally [12].

Polarity, Sentiment Analysis, and Topic Modeling of Data

To measure the polarity values of tweets and get an idea of the underlying sentiments 'Textblob' technique was applied in Python. The polarity ranged from -1 to +1. A polarity score of less than 0 signifies negative sentiments, equal to zero is neutral and more than 0 is the positive sentiment expressed in Tweets [32]. The subjectivity of the tweets was also calculated along with it which reflects the opinionated tweets. R –Studio software was used to analyze emotional tone. 'Tidynext' library was used to classify sentiments. The main emotions were captured based on NRC (National Research Council) word- emotion lexicon association. It consists of 14,182 words in English and calculates not just polarity but also the eight fundamental emotions given by Plutchik [25].

Topic modeling was applied to recognize the most frequent topics evident from the tweets. The probability predicts that the observation belongs to a particular topic. This study uses LDA which is a "generative probabilistic model" comprising of a compilation of composites consisting of parts, which in this study were words or bags of words called n-grams [3]. The two matrices that are formed are:

1. θtd = P(t | d) (the probability distribution of topics in documents)
2. Φwt = P(w | t) (the probability distribution of words in topics) The probability of a word in a given document is:

$$\sum_{t \in T} p(w \mid t, d)\, p(t \mid d)$$

Here T represents total topics. Assuming W is a total number of words that are present in the vocabulary for all documents. Also if we assume Conditional Independence, we have:

P(w | t,d) = P(w | t)

$$\sum_{t=1}^{T} p(w \mid t) p(t \mid d)$$

And hence P(w | d) is equal to:

In the study Topic Model of the text in tweets was done using Python 3.0 software. Genism package is used for the purpose. The data was first pre-processed for topic modeling. This involved three stages, namely:

1. **Tokenization:** The data was converted into sentences and sentences were further converted into words. Also, the words were changed into lowercase, and punctuations, URLs, etc were removed.
2. Stopwords were removed.
3. **Lemmatization:** The words were changed to root form.

Tokenization was done using Spacy, a pre-trained NLP (Natural Language Processing) model that figures out the relationship between words. The perplexity and Coherence score of the model was also calculated.

4. Results

Sentiment Analysis

Polarity and Subjectivity

The polarity of the tweets was calculated first to know the positive, negative or neutral tone of the emotion. To gauge the polarity of the tweets, the number of tweets that were positive, neutral, and negative were plotted as shown in Figure 2.

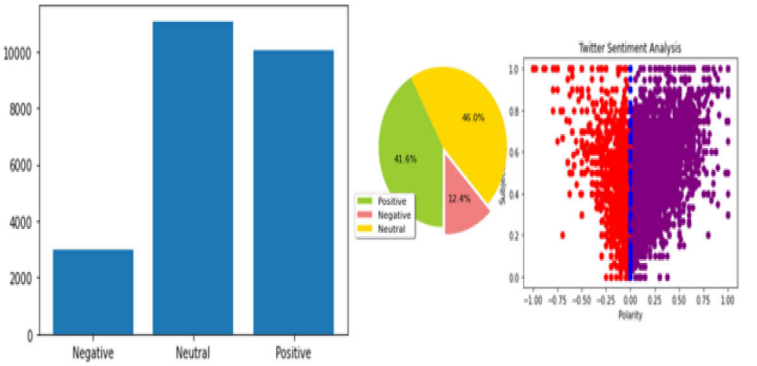

Figure 2

Number of Tweets that were Positive, Negative and Neutral, Pie Chart of percentage of polarity & Scatter plot of Polarity and Subjectivity of the Tweets

It is followed by a pie chart presenting the percentage of the polarity of tweets. It is indicated that the majority of Tweets were neutral (46%) followed by positive tweets (41.6%). The negative tweets were 12.6%.

A Scatter plot of subjectivity and polarity of tweets reflected in Figure 2, indicated that negative (in red color), positive (in purple color), or neutral (in blue color) tweets were subjective or not. Subjectivity ranges from 0 to 1. The majority of the tweets were subjective having a subjectivity value of more than 0.5.

Sentiment Analysis

The sentiment analysis was further enriched by further identifying what emotions are connected with the tweets pertaining to cyber security-related tweets. Figure 3 displays the frequency of the words connected to

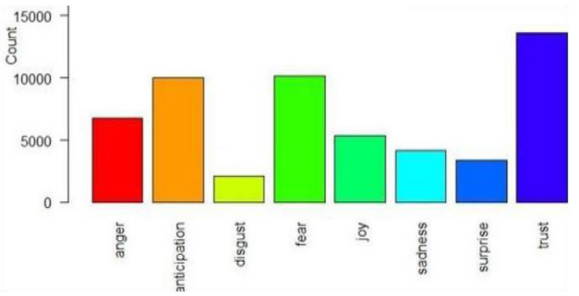

Figure 3

Emotions in the Tweets

a specific emotion of cyber security in the tweets. The emotion of Trust dominated all other emotions followed by that of anticipation and fear.

Topic Modeling

The objective of topic modeling was to know the topical themes discussed in the tweets related to Cyber Security, cybercrime, and cyber attack. The perplexity and coherence values of the topics were found to be 7.246095 and 0.469430. The emergent topics shown in figure 4 are summarized in Table

1. The top 10 words were considered for a topic in the study. The topics latent in the tweets can help organizations and authorities to grasp the themes and concerns of people about Cyber security-related issues.

Table 1

Top Words in Each Theme based on Topic Modeling

Topic Theme	Words
Mode of CyberAttack	Ransomware, malware, hacking, data breach, phishing, data security, cyberattack, infosecurity, threat, privacy
Cyber AttacksImpact	Job, business, government, information security, middlemen, blockchain, hack, intelligence, security
Mitigating strategies	Machine learning, technology, detect, security, python, information technology, informationsecurity

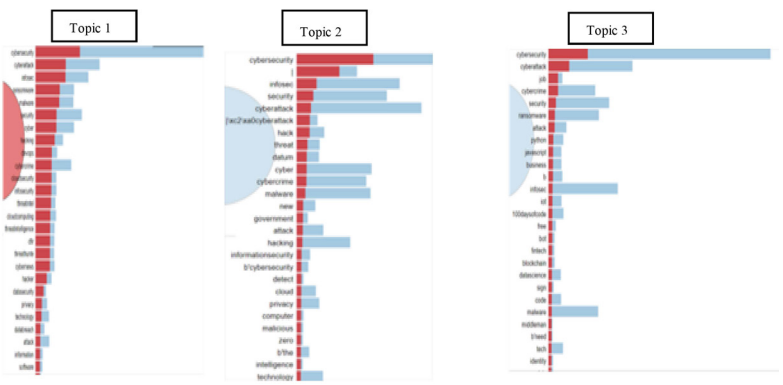

Figure 4

Topic Modelling

5. Discussion and Conclusion

The Twitter data analysis explains the awareness and perception of the users on cybersecurity-related issues. Users, in general, discussed the issues like ransomware, malware, and cyber-attacks in different parts of the world. Since the study took place in November and December 2021, the people were well aware and perhaps more trained to combat the cyber-attacks than initially when the COVID-19 outbreak took place as implied by a large number of positive and neutral tones of the polarity of sentiments The Sentiment Analysis further enlightened that people were having 'trust' as the major emotion in tweets. This may be the people and their organizations have leveraged better tools and techniques to mitigate the online scams. Also, they were more accustomed to working remotely. Moreover, high anticipation emotion also predicts that people were hopeful that the situation will improve and in future, the cyber security will enhance as companies will take better practical approaches to protect their devices and people. Although, the high value of 'Fear" in the sentiment analysis depicts that there are issues people are concerned about and the cyber attacks in Russia, ransomware, hacking, and malware show people are still afraid of cybercrimes. The cyber criminals keep on deploying new techniques to scam the devices and people which require a more precautious take on the issues. But the sentiment analysis reveals as people became more accustomed to the new normal and improved vigilant employment of technology the sentiments were more neutral or positive.

Topic modeling reflected that Twitter users discussed main topics on three main themes the mode of cyber attacks, the impact of cyber attacks, and the mitigating strategies taken by the government and development in data sciences.

There is an up rise in big data analytics in recent times [5]. All policymakers including government and business organizations should recognize the significance of Twitter data in discovering people's awareness and emotions reflected from the tweets. A system management system to safeguard the employees' data and organizational important information should be implemented to safeguard. A more employee-friendly anti cyber- attack system needs to be installed and employees need to be trained to use them. The research has the following limitations which may serve as a guide for future research. The study is based on Twitter users only. Future research may be conducted from other social media and data sources. A study with a broader time frame may be conducted to shed a

line on the transformation of emotions on issues related to cyber-security and mitigation from early stages of pandemic and work from home due to it.

As the pandemic and its impact are not yet over the work from home is new normal where the use of technology is ubiquitous. This calls for short- term and long-term response from organizational authorities as well as government agencies to ensure that cyber-security systems are equipped to handle the escalating cyber-crime cases.

References

[1] Aghajani, G., & Ghadimi, N. (2018). Multi-objective energy management in a micro-grid. *Energy Reports*, *4*, 218-225

[2] Arora, B. (2016). Exploring and analyzing Internet crimes and their behaviours. *Perspectives in Science*, *8*, 540-542.

[3] Singh, A., & Chatterjee, K. (2021). Securing smart healthcare system with edge computing. *Computers & Security*, *108*, 102353.

[4] Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. J. Mach. Learn. Res. 2003,3, 993–1022

[5] Cheung B, Wong CL, Gardhouse A, Frank C, Budd L. #CGS2015: An Evaluation of Twitter Use at the Canadian Geriatrics Society Annual Scientific Meeting. Can Geriatr J 2018 Jun;21(2):166-172

[6] Dadheech, P., Goyal, D., Srivastava, S., & Choudhary, C. M. (2018). An efficient approach for big data processing using spatial Boolean queries. *Journal of Statistics and Management Systems*, *21*(4), 583-591.

[7] Dai H, Deem MJ, Hao J. Geographic variations in electronic cigarette advertisements on Twitter in the United States. Int J Public Health 2017 May;62(4):479-487. [doi: 10.1007/s00038-016-0906-9]

[8] Esuli, A.; Sebastiani, F. SentiWordNet: A high-coverage lexical resource for opinion mining. Evaluation 2007, 17, 26.

[9] Han, G.H.; Jin, S.H. Introduction to big data and the case study of its applications. J. Korean Data Anal. Soc. 2014, 16, 1337–1351.

[10] Han, G.H.; Jin, S.H. Introduction to big data and the case study of its applications. J. Korean Data Anal. Soc. 2014, 16, 1337–1351.

[11] https://www2.deloitte.com/ch/en/pages/risk/articles/impact-covid-cybersecurity.html

[12] Jeong, M.S.; Shon, B.Y. Design and analysis of sentiment classification model for Korean music reviews based on convolutional neural networks. J. Korean Data Anal. Soc. 2018, 20, 1863–1871.

[13] Jivani, A. A Comparative Study of Stemming Algorithms. Int. J. Comp. Tech. Appl. 2011, 2, 1930– 1938.

[14] Judge, M. A., Manzoor, A., Maple, C., Rodrigues, J. J., & ul Islam, S. (2021). Price- based demand response for household load management with interval uncertainty. *Energy Reports*.

[15] Kang, J., Kim, S., & Roh, S. (2019). A topic modeling analysis for online news article comments on nurses' workplace bullying. *Journal of Korean Academy of Nursing*, *49*(6), 736-747.

[16] Kaur, J., & Ramkumar, K. R. (2021). The recent trends in cyber security: a review. *Journal of King Saud University-Computer and Information Sciences*.

[17] Kim, S., Park, H., & Lee, J. (2020). Word2vec-based latent semantic analysis (W2V- LSA) for topic modeling: A study on blockchain technology trend analysis. *Expert Systems with Applications*, *152*, 113401.

[18] Kontopoulos, E., Berberidis, C., Dergiades, T., & Bassiliades, N. (2013). Ontology-based sentiment analysis of twitter posts. *Expert systems with applications*, *40*(10), 4065-4074.

[19] Li, N., Tsigkanos, C., Jin, Z., Hu, Z., & Ghezzi, C. (2020). Early validation of cyber– physical space systems via multi-concerns integration. *Journal of Systems and Software*, *170*, 110742.

[20] Li, Y., & Liu, Q. (2021). A comprehensive review study of cyber-attacks and cyber security; Emerging trends and recent developments. *Energy Reports*, *7*, 8176-8186.

[21] Liu, B. Sentiment Analysis and Opinion Mining. Synth. Lect. Hum. Lang. Technol. 2012, 5, 1– 167.

[22] Martinez-Rojas, M., del Carmen Pardo-Ferreira, M., & Rubio-Romero, J. C. (2018). Twitter as a tool for the management and analysis of emergency situations: A systematic literature review. *International Journal of Information Management*, *43*, 196-208.

[23] Nair MR, Ramya G, Sivakumar PB. Usage and analysis of Twitter during 2015 Chennai flood towards disaster management. Procedia Comp Sci 2017;115:350-358.

[24] Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc* (Vol. 10, No. 2010, pp. 1320-1326).

[25] Parveen, F., Jaafar, N. I., & Ainin, S. (2015). Social media usage and organizational performance: Reflections of Malaysian social media managers. *Telematics and informatics*, *32*(1), 67-78.

[26] Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In *Theories of emotion* (pp. 3-33). Academic press.

[27] Pranggono, B., & Arabo, A. (2021). COVID-19 pandemic cybersecurity issues. *Internet Technology Letters*, *4*(2), e247.

[28] Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-based systems*, *89*, 14-46.

[29] Sarker A, Lakamana S, Hogg-Bremer W, Xie A, Al-Garadi M, Yang Y. Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource. J Am Med Inform Assoc 2020 Aug 01;27(8):1310-1315

[30] Sharma, D., & Khandelwal, S. (2021). Sentiments analysis: Recitation of the concealed emotions of professionals during COVID-19. *Journal of Statistics and Management Systems*, *24*(1), 147- 161.

[31] Sharma, S., & Sharma, S. (2021). Analyzing the depression and suicidal tendencies of people affected by COVID-19's lockdown using sentiment analysis on social networking websites. *Journal of Statistics and Management Systems*, *24*(1), 115-133.

[32] Sidhaye, P., & Cheung, J. C. K. (2015, September). Indicative Tweet Generation: An Extractive Summarization Problem?. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 138-147).

[33] Sohangir, S., Wang, D., Pomeranets, A., & Khoshgoftaar, T. M. (2018). Big Data: Deep Learning for financial sentiment analysis. *Journal of Big Data*, *5*(1), 1-25.

[34] Sun, L., & Yin, Y. (2017). Discovering themes and trends in transportation research using topic modeling. *Transportation Research Part C: Emerging Technologies*, *77*, 49- 66.

[35] Maurya, S., & Jain, A. (2020). Deep learning to combat phishing. *Journal of Statistics and Management Systems*, *23*(6), 945-957.

[36] Tan, S., Xie, P., Guerrero, J. M., Vasquez, J. C., Li, Y., & Guo, X. (2021). Attack detection design for dc microgrid using eigenvalue assignment approach. *Energy Reports*, *7*, 469-476.