Social crisis detection using Twitter based text mining-a machine learning approach

Shoaib Rahman[1], Nusrat Jahan[2,3], Farzana Sadia[1,3], Imran Mahmud[1,4,5]
[1]Department of Software Engineering, Daffodil International University, Dhaka, Bangladesh
[2]Department of Information Technology and Management, Daffodil International University, Dhaka, Bangladesh
[3]Department of Computer Engineering, Universiti Malaysia Perlis, Arau, Malaysia
[4]Graduate School of Business, Universiti Sains Malaysia, Penang, Malaysia
[5]Brac Business School, Brac University, Dhaka, Bangladesh

Article Info

Article history:

Received Apr 24, 2022
Revised Jun 2, 2022
Accepted Sep 30, 2022

Keywords:

Machine learning
Social crisis
Text mining
Twitter

ABSTRACT

Social-media and blogs are increasingly used for social-communication, an idea and thought publishing platform. Public intentions, wisdom, problems, solutions, mental states are shared in social media. Text is being the best and the most common way to communicate over social networks. All kinds of data shared in social sites like Facebook, Twitter, and Microblogs. People from different pursuance uses these media to publish thoughts and convey messages through text. Consequently, occurrences in social life are rapidly discussed in social blogs in daily manner. This work aims at discovering ongoing social crisis from the Twitter data. Text mining technique and sentiment analysis were applied to detect the current social crisis from the social sites. Twitter data were collected to identify the recent social crisis. Furthermore, the identified crisis was compared to reputed newspapers. A hybrid method used to detect recent social issues resulted nicely. However, our proposed analysis shows identifying rate 89%, 95%, 83%, 53%, and 98% for the top 5 identified crisis accordingly in the date between 27 February and 11 March 2020. The strategy used in this study for the detection of recent social crisis will contribute to the social life and findings of crisis will be eliminated easily.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Nusrat Jahan
Department of Information Technology and Management, Daffodil International University
Dhaka, Bangladesh
Email: nusrat.swe@diu.edu.bd

1. INTRODUCTION

The occurrence of social crises nowadays is an influential concern and arduous to eliminate instantly. Crisis in social life becomes a concern of the government and people of a community [1]. Every situation that becomes a reason of danger for social life, people consider it as a social crisis. If it can be identified, bad effects can be minimized and the problem can be eliminated. During any social crisis, reliable information is essential for solving that issue. Many organizations have already utilized the power of social media but a few teams worked for the people in social crisis [2]. Some of the social media and microblogs are popular with users such as Twitter. It is not only the source from where scholars can collect information as a tweet but it is also beneficial for tracking any crisis occurring in social life [3]. A good amount of information can be collected from Twitter to analyze the topic and identify crisis. When something happens in society, the first effect [4] can be seen over social media. Sometimes people do not state anything in real life straight but huge posts, blog writings, and short messages spread over the internet through social media. However, those posts or blog writings can be authentic or fake [5] to make a rumor on social media.

Text mining [6] has taken the challenge to resolve the problem from personal to business reason in this era of internet of things (IoT). When people are nowadays totally depending on social media [7] for news and updates, they are living in virtual life every day. Each problem and solution are posted on social media even though it is not discussed in real life. Text mining can accumulate information that is needed to recognize a problem or to solve the issue. Social media text produces most of the information about what is going on in the world. Natural language processing (NLP) techniques [8] from unstructured data such as Facebook and Twitter text data, topic detection [9] can help by recognizing the most discussed stuff in social life. Sentiment analysis along with text mining [10] has been solving a huge number of issues. It can be used in many different ways to track the target. Different works have been done with sentiment analysis using various lexicon based [11] and machine learning techniques [12] like support vector machine, Naive Bayes, term frequency-inverse document frequency (TF-IDF), unigrams, valence aware dictionary for sentiment reasoning (VADER) by the researchers.

Sentiment analysis [13] is making a great change in this era identifying negative and positive sentiment with reasonable accuracy. From private work to public work or organization everywhere sentiment is important it is being the topic of concern for the scholars [14]. Since people use social media regularly to publish their thoughts and problems, this study's objectives are: i) identify the topic discussed frequently for some days; ii) analyze the sentiment behind the topics; and iii) identify the topic which is a crisis for the specific region and specific time. Twitter was considered to conduct text mining and analyze sentiment for these reasons: i) Twitter is used by millions of users; ii) Twitter is usually used by literate people; iii) Twitter is used all over the world; iv) getting Twitter data is free and easy; v) Twitter data provides age, sex, region, date; and vi) most of the researchers find Twitter data valuable. In this paper, a hybrid method is proposed to identify the recent major social crisis in a specific region and time. Using orange 3, an opensource data mining application [15] this study used the Liu-Hu sentiment analysis method for the following reasons: i) it provides a reasonable accuracy [16] with the dataset of this work; ii) Liu-Hu method in orange is free; iii) Liu-Hu in orange is easy to use for sentiment analysis; iv) it does not need a training dataset to analyze sentiment; and v) proffer a clear view of negative and positive sentiment based on the specific topic. Bag of words model in Spyder is used to classify text and to count word occurrence from social media and to identify the frequently discussed topic for the following reasons: i) bag-of-word is simple to understand; ii) easy to implement; iii) bag-of-word works faster for our dataset; and iv) many scholars prefer bag-of-word for [17] text classification.

Last few years social media text mining and social media emotion mining is a focus point for the researchers. But even so, few fields are existing to work by identifying the gap in the research limitations. From education [18] to tourism [19] social media text [20] and sentiment analysis are being used nowadays to help to make correct decisions [21] clustering the types of short text structures and predicting [22] various factors. Alvermann [23] discussed the critical inquiry in the text of social media. Thelwall used social text to detect magnitude, stress and relaxation [24]. Singh *et al.* [25] talked about the trends ongoing in social media. A customizable pipeline focused by Sarker [26] for social media. Ariffin and Tiun [27] focused on the tagger of parts of speech in his study. According to Pinto *et al.* [28], the performance of NLTK toolkits in social media and regular text is compared. Eryigit and Torunoglu-Selamet [29] discussed the text normalization of social media. According to Hee *et al.* [30], an automated cyberbullying detection carried out. Uteuov and Kalyuzhnaya [31] combined the hierarchical topic model and document embedding for social media. Lexical normalization discussed by Han *et al.* [32]. Wu *et al.* [33] discussed deep learning in sentiment analysis of social media text data. Where the study needs more validation and comparative analysis is possible to interpret online opinions on various microblogs. This study can forward to analyze the sentiment for crisis detection from social media. Mansour [34] studied the thinking of people on social media about the Islamic State of Iraq and Syria (ISIS) using sentiment analysis and text mining. Multilingual tweet analysis and the battle between ISIS and the rest of the world are not done and another topic such as social problem identification can be conducted following this study. In [35], [36] proposed models for sentiment and emotion mining. Suggestion based recommender system can ahead with a larger dataset in this study. Action rules extracted by Ranganathan and Tzacheva [37] concerning the user emotions that help by providing suggestions to enhance users' feelings to lead a better healthy life. Learner's evaluation system for teachers can be implemented with this work. Lexicon-based and machine learning approach used together [38] on product review sentiment study. Emoji is not included in the text to analyze sentiment. Derakhshan and Beigy [39] got better accuracy in stock price movement prediction by sentiment analysis on stock social media. A market simulator can be implemented to estimates profit and lack. A valuable brand sentiment analyzed by Mostafa [40] from social networks to marketing research companies. Most representational topics can be gained and addressed by the work scope that is not done in this study. Only 3,500 tweets considered to conduct the research. According to Canales *et al.* [41], a semiautomatic method for emotion detection from social media text used. More testing and development of new manual jobs with further annotators and a greater amount of the data is needed. According to Chekima and Alfred [42], a process for sentiment analysis from Malay text is conducted. Only Malay text is analyzed. Many other languages can be considered in a new job. News sentiment is measured and generated a sentiment scoring model by Shapiro *et al.* [43] using a sample of articles rated by humans on a scale of positivity and negativity, compared to predictive accuracy at a set of sentiment analysis models, combining existing lexicons and accounting for negation. More accuracy required to improve performance in further work. Comparing support vector machine, Naive Bayes, maximum entropy also the linguistic inquiry and word count (LIWC), the affective norms for english words (ANEW), general inquirer and SentiWordNet. Hutto and Gilbert [44] presented VADER having a great F1 classification accuracy for social media text sentiment analysis. Suppala and Rao [45] measured customer's opinions and perceptions by sentiment analysis on social media data using a naive Bayesian algorithm. The study used the Twitters featured data set, more features in the database can be added to enhance the research.

A chapter on text mining with the unstructured text [46] discussed different techniques on NLP, relation extraction tools, topic modeling, and deep learning. Market prediction using social media text [47] can make strategic decisions. Standardization and comparative performance evaluation are needed in market prediction. Lia *et al.* [48] proposed a realtime monitoring tool for changing customer needs investigation from product planning from social media mining. The approach applied to only one target product where analysis in different disciplines can be done.

There is a lot of research gaps to work with. However, there is no study conducted to identify the current major social concern in recent specific dates and regions. This study targets to identify social issues ongoing from the social media text mining and sentiment analysis [48] to confirm if the target topic is a social crisis or not. This paper concentrates on detecting an ongoing social crisis in an area using both rulebased method and machine learning algorithm. The remainder of this paper is organized as follows: section 2 provides a brief overview of the reviews related literature related to sentiment analysis and text mining. Section 3 presents the architecture of the proposed system. Section 4 discusses how data is collected and the processing of data. Section 5 summarizes conclusions and discusses future work.

2. METHOD

This study developed a method to find out the recent major social crisis using lexical and machine learning techniques. The general strategy is to first target a date from which day we need to identify the crisis. This research targets the last 14 days for the identification of the intricacy. Then identified the frequently used keywords in social media. This work used tweeplers, a trend detector online tool for the tweeter micro-blog posts. From the tweeplers after identification of the common keywords, a search conducted on the recent common keywords in social media (Twitter) with the help of application programmer's interface (API) provided by Twitter. The common keyword indicates the frequently used keywords in user posts on a social site. A region name needs to be added when a country or subcontinent is a target to find out the crisis. Bangladesh was added with the common keywords as the target country to conduct the research. A combination of a region's name and common keywords compels to be added to search tweets. After collection of data set needed to clean and pre-process. Pre-processing techniques of data are discussed in section 4.1. A cleaned dataset is used for topic detection with the model bag-of-words. Bag-of-word performed surprisingly well in topic identification in this study. The technique is described in section 3.2. When topics are identified, the study cannot determine which are the major social concern yet. In the next step classified and prioritized topics are then used for sentiment analysis. Prioritized topics are categorized into two sections, negative and positive to determine the crisis. This study used the Liu hu model for sentiment analysis discussed in section 3.3. Later on, analyzing the sentiment on the topics, the result is compared with some of the top reliable online regional newspapers of the specific country, which is considered to recognize the major social crisis between those days. Crisis identification architecture is described in Figure 1.

2.1. Text mining based topic identification

Bag of words a decent machine learning technique, used for Twitter text mining. This study used bag-of-words to identify the word calculation from the tweets. The natural language toolkit (NLTK) tokenization is used first to tokenize sentences. String substitution used to normalize whitespace and to extract hypertext markup language (HTML) tags. Each word transformed into lower text before lemmatization and the creation of a corpus. A corpus is used to count repeatedly used words increasing value by 1. An algorithm 1 to count repeated words in a dataset is stated as (1).

Here major issue holds the value of each word count for each of the words. From the tweets gathered From Twitter, each word is counted uniquely. When a word is reproduced, word count increases by value +1, or it keeps its value as the earlier value. The initial value for each word is 0. Executing the same manner for the whole dataset, each word is computed. The highest used keyword in a dataset holds the maximum value. Based on word count descending order is ensued to prioritize subjects' array. A representation of prioritized tweets data is shown in Table 1. Here the maximum used keyword is on the top of the list then the second one and so on. Top 10 prioritized topics from this table are practiced in this research to identify the major issues in social networking micro-blog site, Twitter. Some unwanted and unnecessary data can be seen in the table that creates no sensible value. Data is again concocted for topic classification to avoid outcast and unnecessary topics, addressed in data collection and processing section 4.1. After excluding extraneous words that do not create sentiment value e.g. country name, place name a precise output is shown in Table 2 as the identified topics.

Algorithm 1

```
Major Issue[Word]
If(word not in MajorIssue) : MajorIssue[word] = 1
Else:MajorIssue[word]+=1                                  (1)
```

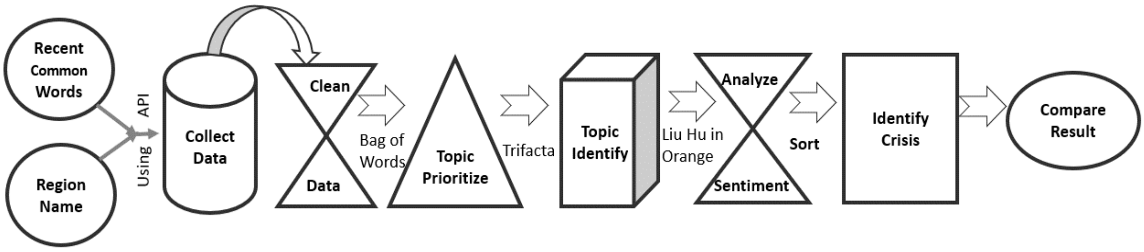

Figure 1. Crisis identification architecture

Table 1. Prioritized topics

Key	Type	Size	Value
Modi	int	1	6,534
Pakistan	Int	1	4,997
Coronavirus	Int	1	4,846
Hindu	Int	1	4,720
March	Int	1	3,657
Temple	Int	1	3,037
Dhaka	Int	1	2,772
Place	Int	1	2,608
Islamist	int	1	2,517

Table 2. Identified topics

Key	Type	Size	Value
Modi	Int	1	4,997
coronavirus	int	1	4,720
Hindu	Int	1	3,657
March	Int	1	3,037
Temple	Int	1	2,772
Dhaka	Int	1	2,757
islamist	int	1	2,717

2.2. Text mining based topic identification

Sentiment analysis is used to analyze the text sentiment. Some studies have appropriated sentiment analysis techniques in crisis domain for detecting the sentiments of posts on disaster management. Sentiment analysis is used to identify the sentiment of classified prioritized topics. This study used a lexicon based sentiment analysis model Liu-Hu in orange. Orange is a desktop-based data visualization and analysis tool used by many researchers, which is opensource. Liu-Hu uses sentiment modules from NLTK. Two types of results are considered here for our dataset, positive sentiment, and negative sentiment. If the topic relevant sentences proffer maximum positive value, it is acknowledged as a exclude topic. The negative value is worthy here to identify the crisis. Figure 2 and Table 3 draw the sentiment analysis from the classified topics. Here all the identified topics are chosen to analyze sentiment from the text. Figure 2 is for each of the topics is analyzed in this way in the orange tool using the Liu-Hu method. It provides -10 to +10 sentiment value for each sentence.

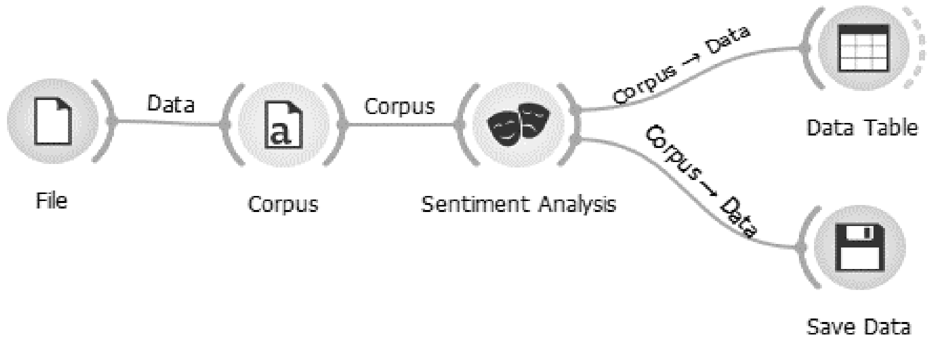

Figure 2. Sentiment analysis using Liu-Hu in orange

The total sentence count in the dataset related to a specific topic and positive sentences that have no preference in this study. This makes no quandary or not a topic to be a crisis, the possible crisis is no here. In the same manner, the neutral sentences make no sense whether it is positive or negative nor a crisis. It gives straight 0 value output, which is considered as a positive number and a possible crisis is measured as no. The negative sentences attest that the topic can be the crisis for a region at a specific type. Average sentiment value is the focal point here to determine the crisis showed in Table 3. Once the average sentiment value is calculated, positive and neutral topics are eliminated in Table 4. From the existing negative sentiment topics, the study determines the major issue as a rank column based on topic frequency count. The most frequently discussed topic with negative sentiment value is determined as a crisis rank 1 in this study. Then the following next topic and so on, stated in Table 5. Excluding the positive sentiment value, average sentiment shows the effect of public posts in this study.

Table 3. Identified possible crisis

Date between	Keyword	Total data	kW used	Positive data	Neutral	Negative data	Avg sentiment	Possible crisis
27/2/2020 to 11/3/2020	Dhaka	28,264	27,867	4,932	15,670	7,265	0.74961	Yes
27/2/2020 to 11/3/2020	Modi	28,264	6,767	1,044	2,937	2,786	-2.475996972	Yes
27/2/2020 to 11/3/2020	Coronavirus	28,264	4,351	342	1,923	2,086	-2.10068	Yes
27/2/2020 to 11/3/2020	Hindu	28,264	3,476	149	543	2,784	-5.23379	Yes
27/2/2020 to 11/3/2020	Temple	28,264	2,762	47	488	2,227	-3.0303	Yes
27/2/2020 to 11/3/2020	Islamist	28,264	2,516	38	214	2,264	-6.5698	Yes
27/2/2020 to 11/3/2020	March	28,264	1,664	708	706	250	0.716042	No

Table 4. Eliminated positive and neutral topics

Date between	Keyword	Total data	kW used	Positive data	Neutral	Negative data	Avg sentiment	kW rank
2020-02-27-2020-03-11	Modi	28,264	6,767	1,022	2,937	2,786	-2.475996972	1
2020-02-27-2020-03-11	Coronavirus	28,264	4,351	342	1,923	2,086	-2.10068	2
2020-02-27-2020-03-11	Hindu	28,264	3,476	149	543	2,784	-5.23379	3
2020-02-27-2020-03-11	Temple	28,264	2,762	47	488	2,227	-3.0303	4
2020-02-27-2020-03-11	Islamist	28,264	2,516	38	214	2,264	-6.5698	5

Table 5. Identified major crisis sentiment level

Date between	Keyword	Total data	kW used	Positive data	Neutral	Negative data	Avg sentiment	kW rank
2020-02-27-2020-03-11	Islamist	28,264	2,516	38	214	2,264	-6.5698	1
2020-02-27-2020-03-11	Hindu	28,264	3,476	149	543	2,784	-5.23379	2
2020-02-27-2020-03-11	Temple	28,264	2,762	47	488	2,227	-3.0303	3
2020-02-27-2020-03-11	Modi	28,264	6,767	1,022	2,937	2,786	-2.475996972	4
2020-02-27-2020-03-11	Coronavirus	28,264	4,351	342	1,923	2,086	-2.10068	5

3. RESULTS AND DISCUSSION

3.1. Data collection and processing

The initial query for data collection performed from 11th February 2020 to 11th March 2020 on Twitter. An API was used for tweets collection provided by Twitter to a developer account. The keywords used were “crisis”, “attention”, “Dhaka”, “virus”, “issue”, “religion”, “coronavirus”, “want justice”, “problem” to retrieve data from the micro-blog. “Bangladesh” and “Dhaka” were used in addition to these keywords to search the data. Here Bangladesh and Dhaka were applied as the target location to identify the crisis in march 2020. Till the last day, 28,264 tweets were collected using the specific keywords for the region Bangladesh. After that, we performed data processing to prepare our data. Trifacta, an online data wrangler tool was used to get cleaned and structured data formats. This study used some techniques to prepare the data through Trifacta are the following: i) removed unwanted columns; ii) trimmed whitespaces; iii) removed URLs; iv) removed symbols; v) removed accents from texts; vi) removed hash (#) from the hashtags; vii) removed usernames from the text; viii) removed retweets; and ix) removed at (@) sign from posts. Then, the processed data was used to identify the social crisis for Bangladesh.

3.2. Evaluation

The main goal of this study was to identify the recent social crisis from the social media text. To evaluate the method, the study conducted two appraisals. First collected some reliable newspapers from Bangladesh which were published after the date of 11th March 2020 to 15th March 2020. The study searched topics in 11 newspapers randomly including the daily star, daily observer, Dhaka Tribune, daily sun, Bangladesh today, financial express, and prothom-alo. Then the study focused on the topics of the newspapers those headlines matched with our identified topics. If it is matched then, after a manual reading, we numbered it positive and negative based on the effect of those topics discussed in newspapers. We considered financial, social, religious, and political effects in this job. When the topic was not clear to demonstrate the core idea to score the value it was marked as zero or neutral. Also, topics that were not present in the newspapers were scored as neutral. After analyzing data of the next 4 days from the data collection date, we found the effect of those topics on social life. The total paper citation was 44 times including all the papers. Negative scores of all the topics create value here. After getting the output from the newspapers we compared the result with our identified topics. The compared result is stated in Figure 3. The chart shows the comparison of the study result and the newspaper result. It seemed the results matched the prediction of this research. Finally, the result of this study could predict the correct output as the crisis depending on the word frequency and sentiment analysis. The sentiment score was also matched wonderfully with the sentiment level of the local public news described in newspapers. However, this research focused on the recent crisis, identifying the negative and positive sentiment value.

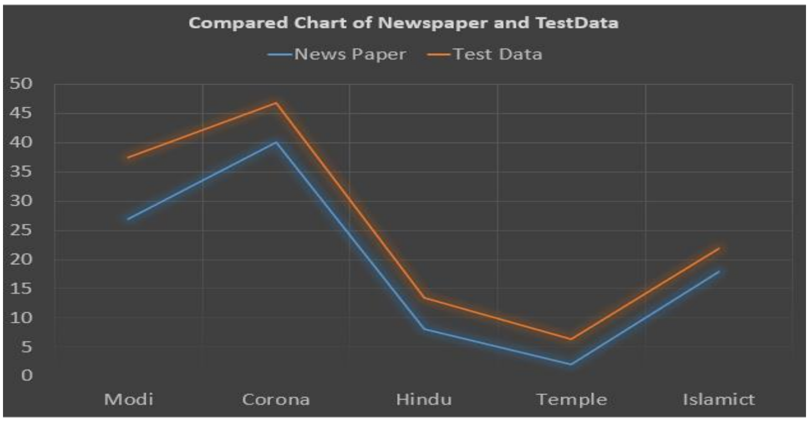

Figure 3. Comparison chart with newspaper

From Table 6, we can observe, the identified confusion matrix of collected and verified data. Correctly identification rate is 89%, 95%, 83%, 53%, and 98% respectively for the words ‘Modi’, ’Corona’, ’Hindu’, ’Temple’ and ’Islamist’. 53% for one topic was considered comparing with newspapers because some topics never get picked up by the news media for many reasons but those are absolutely crisis making and more discussed topics in social media sometimes. Frequently used negative words were taken as the test data and newspaper data as true data. Rate over 0.50% was taken as 1. The total keywords count was calculated from the data and subtracted the natural values. After subtraction negative or positive frequent keywords count, rest words were divided and the result was multiplied by 100 to get the rate of negative and positive value. The used formulas for identifying cases were:

$$S = (nkw/(kw - n)) * 100 \tag{2}$$

$$S = (pkw/(kw - n)) * 100 \quad (3)$$

In (2) used for identifying negative keywords from test data and newspaper data also (3) used for positive keywords. Where S is the result and nkw means negative keywords, kw means total keywords of the specific topic, n means natural words. This study identified true-positive, false-positive, true-negative, and false-negative data. Table 1 shows the identifying rate of data.

Table 6. Confusion matrix

	TN	TP	FN	FP
Modi	73	16	0	11
Corona	86	9	0	5
Hindu	80	5	15	0
Temple	50	3	47	0
Islamist	98	0	0	2

3.3. Limitation and future work

Due to changes in the organizational structure of Twitter, there might be an effect on users. In future we may work on, social crisis during COVID-19 [49] period by using Twitter data and may find their probable intentions [50]. Again, combined model of machine leaning and non-machine learning models can be applied in further development of this work [51].

4. CONCLUSION

A crisis detection approach was proposed in this study for identifying the ongoing social problems. Dates between 27 February 2020 and 11 March 2020 was selected to collect data to determine the crisis at that time. Local newspapers were used as the validator of the detected crisis to measure the result of the output we get from this study. Bangladesh was used as the location for the crisis zone. This study used the machine learning-based bag of-words method and the lexicon based Liu-Hu sentiment analysis method, which worked great to conduct the result with the collected dataset. As a later form of this work, it is important to introduce an algorithm that can offer a decision support system from the result of this work as a recognized social crisis. A decision support system along with this work can eliminate the crisis or can reduce the effect of the crisis in social life. We expect that the study will determine the crisis from social media and it will assist to reduce the problems to help both the public and the government of any specific region.

REFERENCES

[1] P. Panagiotopoulos, J. Barnett, A. Z. Bigdeli, and S. Sams, “Social media in emergency management: Twitter as a tool for communicating risks to the public,” *Technological Forecasting and Social Change*, vol. 111, pp. 86–96, Oct. 2016, doi: 10.1016/j.techfore.2016.06.010.

[2] L. Austin and Y. Jin, *social media and crisis communication*. New York: Routledge, 2017, doi: 10.4324/9781315749068.

[3] O. D. Apuke and E. A. Tunca, “Social media and crisis management: A review and analysis of existing studies,” *LAÜ Sosyal Bilimler Dergisi*, vol. 9, no. 2, pp. 199–215, 2018.

[4] N. A. B. Ibrahim, R. B. Musa, and R. B. A. Wahab, “The effect of social media depends on social media intelligence among graduates,” in *Regional Conference on Science, Technology and Social Sciences (RCSTSS 2014)*, Singapore: Springer Singapore, 2016, pp. 835–843, doi: 10.1007/978-981-10-1458-1_76.

[5] M. Aldwairi and A. Alwahedi, “Detecting fake news in social media networks,” *Procedia Computer Science*, vol. 141, pp. 215–222, 2018, doi: 10.1016/j.procs.2018.10.171.

[6] M. Pejic-Bach, T. Bertoncel, M. Meško, and Ž. Krstić, “Text mining of industry 4.0 job advertisements,” *International Journal of Information Management*, vol. 50, pp. 416–431, Feb. 2020, doi: 10.1016/j.ijinfomgt.2019.07.014.

[7] O. Berezan, A. S. Krishen, S. Agarwal, and P. Kachroo, “The pursuit of virtual happiness: Exploring the social media experience across generations,” *Journal of Business Research*, vol. 89, pp. 455–461, Aug. 2018, doi: 10.1016/j.jbusres.2017.11.038.

[8] J. Bhattacharjee, *Practical machine learning with Rust: Creating intelligent applications in Rust*. Bangalore: Apress, 2019, doi: 10.1007/9781484251218.

[9] D. Milioris, *Topic detection and classification in social networks*. Cambridge: Springer, 2017.

[10] C. A. Iglesias and A. Moreno, “Sentiment analysis for social media,” *Applied Sciences*, vol. 9, no. 23, pp. 1–4, Nov. 2019, doi: 10.3390/app9235037.

[11] D. Ray, “Lexicon based sentiment analysis of Twitter data,” *International Journal for Research in Applied Science and Engineering Technology*, vol. 5, no. 10, pp. 910–915, Oct. 2017, doi: 10.22214/ijraset.2017.10130.

[12] S. Redhu, “Sentiment analysis using text mining: A review,” *International Journal on Data Science and Technology*, vol. 4, no. 2, pp. 49–53, 2018, doi: 10.11648/j.ijdst.20180402.12.

[13] J. Shen, O. Baysal, and M. O. Shafiq, “Evaluating the performance of machine learning sentiment analysis algorithms in software engineering,” in *2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress*

(DASC/PiCom/CBDCom/CyberSciTech), Aug. 2019, pp. 1023–1030, doi: 10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00185.

[14] T. Nguyen, D. Phung, B. Adams, and S. Venkatesh, "Mood sensing from social media texts and its applications," *Knowledge and Information Systems*, vol. 39, no. 3, pp. 667–702, Jun. 2014, doi: 10.1007/s10115-013-0628-8.

[15] Biolab, "Orange3 text mining documentation," *Readthedocs.org*, 2020. https://readthedocs.org/projects/orange3-text/downloads/pdf/latest/ (accessed Mar. 17, 2020).

[16] S. Kolevska, "Data mining using orange - interworks," *InterWorks*, 2020. https://interworks.com.mk/data-mining-using-orange/ (accessed Mar. 26, 2020).

[17] A. Kadriu, L. Abazi, and H. Abazi, "Albanian text classification: Bag of words model and word analogies," *Business Systems Research Journal*, vol. 10, no. 1, pp. 74–87, Apr. 2019, doi: 10.2478/bsrj-2019-0006.

[18] R. Ferreira-Mello, M. André, A. Pinheiro, E. Costa, and C. Romero, "Text mining in education," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 6, pp. 1–49, Nov. 2019, doi: 10.1002/widm.1332.

[19] Y. Tao, F. Zhang, C. Shi, and Y. Chen, "Social media data-based sentiment analysis of tourists' air quality perceptions," *Sustainability*, vol. 11, no. 18, pp. 1–23, Sep. 2019, doi: 10.3390/su11185070.

[20] K. K. Kapoor, K. Tamilmani, N. P. Rana, P. Patil, Y. K. Dwivedi, and S. Nerur, "Advances in social media research: Past, present and future," *Information Systems Frontiers*, vol. 20, no. 3, pp. 531–558, Jun. 2018, doi: 10.1007/s10796-017-9810-y.

[21] C. E. H. Chua, V. C. Storey, X. Li, and M. Kaul, "Developing insights from social media using semantic lexical chains to mine short text structures," *Decision Support Systems*, vol. 127, pp. 1–10, Dec. 2019, doi: 10.1016/j.dss.2019.113142.

[22] T. Nguyen, D. Phung, B. Adams, and S. Venkatesh, "Prediction of age, sentiment, and connectivity from social media text," *International Conference on Web Information Systems Engineering*. pp. 227–240, 2011, doi: 10.1007/978-3-642-24434-6_17.

[23] D. E. Alvermann, "Social media texts and critical inquiry in a post-factual era," *Journal of Adolescent and Adult Literacy*, vol. 61, no. 3, pp. 335–338, 2017, doi: 10.1002/jaal.694.

[24] M. Thelwall, "TensiStrength: Stress and relaxation magnitude detection for social media texts," *Information Processing & Management*, vol. 53, no. 1, pp. 106–121, Jan. 2017, doi: 10.1016/j.ipm.2016.06.009.

[25] T. Singh, M. Kumari, T. L. Pal, and A. Chauhan, "Current trends in text mining for social media," *International Journal of Grid and Distributed Computing*, vol. 10, no. 6, pp. 11–28, Jun. 2017, doi: 10.14257/ijgdc.2017.10.6.02.

[26] A. Sarker, "A customizable pipeline for social media text normalization," *Social Network Analysis and Mining*, vol. 7, no. 1, pp. 1–13, Dec. 2017, doi: 10.1007/s13278-017-0464-z.

[27] S. N. A. N. Ariffin and S. Tiun, "Part-of-speech tagger for malay social media texts," *GEMA Online Journal of Language Studies*, vol. 18, no. 4, pp. 124–142, Nov. 2018, doi: 10.17576/gema-2018-1804-09.

[28] A. Pinto, H. G. Oliveira, and A. O. Alves, "Comparing the performance of different NLP toolkits in formal and social media text," *OpenAccess Series in Informatics*, vol. 51, pp. 31–316, 2016, doi: 10.4230/OASIcs.SLATE.2016.3.

[29] G. Eryiğit and D. Torunoğlu-Selamet, "Social media text normalization for Turkish," *Natural Language Engineering*, vol. 23, no. 6, pp. 835–875, 2017, doi: 10.1017/S1351324917000134.

[30] C. van Hee *et al.*, "Automatic detection of cyberbullying in social media text," *PLOS ONE*, vol. 13, no. 10, pp. 1–22, Oct. 2018, doi: 10.1371/journal.pone.0203794.

[31] A. Uteuov and A. Kalyuzhnaya, "Combined document embedding and hierarchical topic model for social media texts analysis," *Procedia Computer Science*, vol. 136, pp. 293–303, 2018, doi: 10.1016/j.procs.2018.08.285.

[32] B. Han, P. Cook, and T. Baldwin, "Lexical normalization for social media text," *ACM Transactions on Intelligent Systems and Technology*, vol. 4, no. 1, pp. 1–27, Jan. 2013, doi: 10.1145/2414425.2414430.

[33] P. Wu, X. Li, S. Shen, and D. He, "Social media opinion summarization using emotion cognition and convolutional neural networks," *International Journal of Information Management*, vol. 51, pp. 1–15, Apr. 2020, doi: 10.1016/j.ijinfomgt.2019.07.004.

[34] S. Mansour, "Social media analysis of user's responses to terrorism using sentiment analysis and text mining," *Procedia Computer Science*, vol. 140, pp. 95–103, 2018, doi: 10.1016/j.procs.2018.10.297.

[35] J. K. Rout, K.-K. R. Choo, A. K. Dash, S. Bakshi, S. K. Jena, and K. L. Williams, "A model for sentiment and emotion analysis of unstructured social media text," *Electronic Commerce Research*, vol. 18, no. 1, pp. 181–199, Mar. 2018, doi: 10.1007/s10660-017-9257-8.

[36] K. Sailunaz and R. Alhajj, "Emotion and sentiment analysis from Twitter text," *Journal of Computational Science*, vol. 36, pp. 1–18, Sep. 2019, doi: 10.1016/j.jocs.2019.05.009.

[37] J. Ranganathan and A. Tzacheva, "Emotion mining in social media data," *Procedia Computer Science*, vol. 159, pp. 58–66, 2019, doi: 10.1016/j.procs.2019.09.160.

[38] A. Veluchamy, H. Nguyen, M. L. Diop, and R. Iqbal, "Comparative study of sentiment analysis with product reviews using machine learning and lexicon-based approaches," *SMU Data Science Review*, vol. 1, no. 4, pp. 1–22, 2018.

[39] A. Derakhshan and H. Beigy, "Sentiment analysis on stock social media for stock price movement prediction," *Engineering Applications of Artificial Intelligence*, vol. 85, pp. 569–578, Oct. 2019, doi: 10.1016/j.engappai.2019.07.002.

[40] M. M. Mostafa, "More than words: Social networks' text mining for consumer brand sentiments," *Expert Systems with Applications*, vol. 40, no. 10, pp. 4241–4251, Aug. 2013, doi: 10.1016/j.eswa.2013.01.019.

[41] L. Canales, W. Daelemans, E. Boldrini, and P. Martinez-Barco, "EmoLabel: Semi-automatic methodology for emotion annotation of social media text," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 579–591, Apr. 2022, doi: 10.1109/TAFFC.2019.2927564.

[42] K. Chekima and R. Alfred, "Sentiment analysis of Malay social media text," *International Conference on Computational Science and Technology*. pp. 205–219, 2018, doi: 10.1007/978-981-10-8276-4_20.

[43] A. H. Shapiro, M. Sudhof, and D. J. Wilson, "Measuring news sentiment," *Journal of Econometrics*, vol. 228, no. 2, pp. 221-243, Mar. 2020, doi: 10.24148/wp2017-01.

[44] C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, 2014, pp. 216–225.

[45] K. Suppala and N. Rao, "Sentiment analysis using naïve bayes classifier," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 8, no. 8, pp. 265–269, 2019.

[46] C. Chen and M. Song, "Text mining with unstructured text," in *Representing Scientific Knowledge*, Cham: Springer International Publishing, 2017, pp. 223–261, doi: 10.1007/978-3-319-62543-0_6.

[47] A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, and D. C. L. Ngo, "Text mining for market prediction: A systematic review," *Expert Systems with Applications*, vol. 41, no. 16, pp. 7653–7670, Nov. 2014, doi: 10.1016/j.eswa.2014.06.009.

[48] R. J. Lia, A. B. Siddikk, F. Muntasir, S. S. M. M. Rahman, and N. Jahan, "Depression detection from social media using Twitter's tweet," in *Big Data Intelligence for Smart Applications*, Cham: Springer, 2022, pp. 209–226, doi: 10.1007/978-3-030-87954-9_9.

[49] E.-U. Rahaman, I. Mahmud, M. A. R. Himel, A. Begum, and N. Jahan, “Mathematical modelling of teachers’ intention to participate in online training during COVID-19 lockdown: Evidence from emerging economy,” *International Journal of Emerging Technologies in Learning (iJET)*, vol. 17, no. 12, pp. 170–183, Jun. 2022, doi: 10.3991/ijet.v17i12.30465.

[50] N. Jahan, M. A. H. Shawon, F. Sadia, D. K. Nitu, M. E. K. Ribon, and I. Mahmud, “Modelling consumer’s intention to use iot devices: role of technophilia,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 23, no. 1, pp. 612–620, Jul. 2021, doi: 10.11591/ijeecs.v23.i1.pp612-620.

[51] A. Rahman, T. Ai Ping, S. K. Mubeen, I. Mahmud, and G. A. Abbasi, "What influences home gardeners’ food waste composting intention in high-rise buildings in Dhaka Megacity, Bangladesh? An integrated model of TPB and DMP," *Sustainability*, vol. 14 no 15, p. 9400, 2022, doi: 10.3390/su14159400.

BIOGRAPHIES OF AUTHORS

Shoaib Rahman is working as a Data Engineer, Team Lead at V2 Technologies Ltd. and Data Engineer Instructor at AIQuest, Bangladesh. He completed his B.Sc. in Software Engineering and finishing M.Sc in Software Engineering from Daffodil International University. He is interested in big data technology management, data engineering, database engineering, machine learning and artificial intelligence. He can be contacted at email: shoaib35-1688@diu.edu.bd.

Nusrat Jahan is working as an Assistant Professor and Head at department of Information Technology & Management in Daffodil International University, Bangladesh. She completed her M.Sc. and B.Sc. in Information Technology from Institute of Information Technology, Jahangirnagar University. She is doing her PhD from department of computer engineering, University Malaysia Perlis (UniMap). She is interested in technology management, computer networks, machine learning and artificial intelligence. She can be contacted at email: nusrat.swe@diu.edu.bd.

Farzana Sadia received is working as an Assistant Professor at department of Software Engineering in Daffodil International University in Bangladesh. She completed her M.Sc in Software Engineering from Independent University, Bangladesh and B.Sc in Computer Science and Engineering, Ahsanullah University of Science and Technology. She is interested in technology management, computer networks, machine learning and artificial intelligence. She can be contacted at email: sadia.swe@diu.edu.bd.

Dr. Imran Mahmud received is an associate professor and head at Department of Software Engineering at Daffodil International University. He is also an associate director of research. Dr. Imran completed his PhD in technology management from Universiti Sains Malaysia. His research interests are human computer interaction, usability testing, software engineering measurement/models and management information systems. Dr. Imran has several research papers on enterprise resource planning and information system published in Sage and IEEE. He can be contacted at email: imranmahmud@daffodilvarsity.edu.bd.