Web mining and sentiment analysis of COVID-19 discourse in online forum communities

Masurah Mohamad[1], Suraya Masrom[1], Khairulliza Ahmad Salleh[1], Lathifah Alfat[2], Muhammad Nasucha[2], Nur Uddin[2]
[1]Computing Sciences Studies, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA Perak Branch, Tapah Campus, Malaysia
[2]Department of Informatics, Universitas Pembangunan Jaya, South Tangerang, Indonesia

Article Info

Article history:

Received Nov 20, 2023
Revised Feb 16, 2024
Accepted Feb 20, 2024

Keywords:

COVID-19
Data scientists
Machine learning
Sentiment analysis
Web mining

ABSTRACT

Recently, various discussions, solutions, data, and methods related to coronavirus disease 2019 (COVID-19) have been posted in online forum communities. Although a vast amount of posting on COVID-19 analytical projects are available in the online forum communities, much of them remain untapped due to limited overview and profiling that focuses on COVID-19 analytic techniques. Thus, it is quite challenging for information diggers and researchers to distinguish the recent trends and challenges of COVID-19 analytic for initiating different and critical studies to fight against the coronavirus. This paper presents the findings of a study that executed a web mining process on COVID-19 data analytical projects from the Stack Overflow and GitHub online community platforms for data scientists. This study provides an insight on what activities can be conducted by novice researchers and others who are interested in data analysis, especially in sentiment analysis. The classification results via Naïve Bayes (NB), support vector machine (SVM) and logistic regression (LR) have returned high accuracy, indicating that the constructed model is efficient in classifying the sentiment data of COVID-19. The findings reported in this paper not only enhance the understanding of COVID-19 related content and analysis but also provides promising framework that can be applied in diverse contexts and domains.

Corresponding Author:

Masurah Mohamad
Computing Sciences Studies, College of Computing, Informatics and Mathematics
Universiti Teknologi MARA Perak Branch, Tapah Campus
Perak, Malaysia
Email: masur480@uitm.edu.my

1. INTRODUCTION

The coronavirus disease 2019 (COVID-19) pandemic has transformed almost all our daily activities such as education activities [1], food-related activities [2] and economic activities [3] into a new normal. Not only daily activities, research activities from different fields have also moved towards COVID-19 related issues. Data scientists have undertaken numerous initiatives to develop pertinence software with different computing techniques to analyze, visualize, track, predict, forecast, and potentially alleviate the COVID-19 phenomenon. The ultimate goal from these works is to provide more insights that will then trigger new analytical activities by data scientists and the development of software tools, specifically tailored to the COVID-19 and potentially reshaping future practices when the same or new pandemics occur.

The most significant technique used by the data scientists to cater to big data evolution on COVID-19 is web mining [4]. Latif *et al.* [5] have presented bibliometric analysis based on web mining dataset on COVID-19 spread and mitigation strategies. The findings present relevant use cases of data science activities and summarizing publicly available datasets to be used by researchers. Due to emerging challenges related to COVID-19 software tools, researchers in [6]–[8] have utilized web mining to get insights on the recent trends of software development to combat COVID-19. Furthermore, in concern to software development challenges in handling bugs and errors, there were 129 open-source software projects for COVID-19 in GitHub that have been analyzed, and the researchers have successfully classify 550 bugs based on the proposed machine learning prediction project [9]. GitHub is a web-based collaboration platform for software developers and has been used as a popular platform for web mining including in [10], which were the first two projects related to COVID-19. Stack Overflow is another platform for data scientists and software developers. Georgiou *et al.* [11], the Stack Overflow platform has been analyzed with web mining technique to present the most prominent technologies adopted for developing COVID-19 software. The results have shown that developers community response were immediate and the interest of developers on COVID-19 related challenges was sustained after its initial peak.

Beyond web or data mining, sentiment analysis which also known as opinion mining (OA) emerges as a valuable technique for examining and identifying the emotional tone or attitude conveyed in various forms of communication, including text, speech, and other types of media [12]. The majority of COVID-19 sentiment analyses have predominantly utilized social media platforms, notably Twitter [13]–[15]. Sentiment analysis faces challenges such as categorizing text in the form of sarcasm, irony, context-dependent sentiment, and handling mixed sentiments within a single text. Businesses use it to analyze customer feedback, reviews, and social media comments to gauge customer satisfaction [16], [17]. The problem arises from the limited focus on popular social media platforms, such as Twitter, for sentiment analysis and data mining, which might not provide a holistic representation of the diverse approaches and discussions within the data science community. This study addresses the gap and aims to rectify it by looking into GitHub and Stack Overflow platforms, where data scientists and developers actively collaborate on COVID-19 analytical projects.

The contributions of this study are threefold. First, this study contributes by preparing a comprehensive COVID-19 dataset obtained through meticulous web mining activities on GitHub and Stack Overflow platforms. The dataset is a rich resource encompassing diverse perspectives and practices from the global data science community engaged in COVID-19 analytical projects. Second, the study introduces a robust research framework based on natural language processing (NLP) and machine learning analyses. Since recent literatures do not provide detail implementation framework for analyzing COVID-19 dataset based on NLP and machine learning, this framework that was constructed based the prepared COVID-19 dataset, provides a structured approach for extracting meaningful insights, sentiments, and patterns from textual data. Thus, this will contribute to the advancement of analytical methodologies in the context of the ongoing pandemic. It will enable researchers and practitioners to consistently use and build the proposed framework, ensuring reliable and comparable analyses in the field. Third, the study enhances analytical depth by incorporating semantic analysis within COVID-19-related content. This leads to a more comprehensive interpretation of the data, offering valuable insights into the complexities of the pandemic issues for future intervention strategies.

This paper comprises of four sections. The first section provided an overview of the study's background. In the second section, the methodology used to conduct the study was described, including the flow of the process. The third section explained the experiments that were carried out, along with the results obtained. Finally, the fourth section drew conclusions based on the overall findings of the study.

2. METHOD

This study involved important methods for achieving the objectives, including sentiment analysis, and machine learning. Machine learning for classification is a powerful application where algorithms are trained to categorize or classify input data into predefined classes or labels [18]. Three algorithms were utilized for the machine learning analysis namely linear regression, Naïve Bayes (NB) and support vector machine (SVM). Linear regression is widely used for predicting numerical outcomes and understanding the strength and direction of relationships between variables [19], [20]. NB is a probabilistic algorithm that often performs well and is particularly popular for text classification tasks, spam filtering, and sentiment analysis [21]. It is a relatively simple yet powerful algorithm that works well in practice, especially when the independence assumption holds reasonably well or when computational efficiency is a priority [22]. SVM is a powerful supervised machine learning algorithm that is used for both classification and regression tasks. SVM are widely used in various domains due to their effectiveness, especially in high-dimensional spaces such as in image classification and text classification [23].

2.1. Research phases

This study consists of four main phases, including data collection, feature selection and extraction, data analysis, and results generation and visualization. The proposed work process is depicted in Figure 1. To gather the desired data, specific queries were employed to locate and save the targeted data in the required format. Several tools and application programming interfaces (APIs) can be used to scrape data from websites such as beautiful soap and regular expressions. These three examples are Python packages and library that helps user to collect data from desired pages [24]. The data mining process utilized web scraping methods to visit various web forums and determine which communities provided sufficient data for collection and analysis. General keywords were used to collect more data for analysis. The data were then processed through feature selection and extraction phases, which included stop word removal and word stemming processes to further clean and prepare the data. URL links, white spaces, symbols, abbreviations, and other extraneous information are several items that will be removed from the raw collected dataset [25].

Next process is data analysis that started with word vectorization (transforming the word to numerical presentation) [26], and identifying word polarity using TextBlob [17]. TextBlob is a lexicon-based text analyzer which applies natural language toolkit (NLTK) to process the text [17]. Word polarity is used in identifying the score of the word to either positive, negative, or neutral. If the polarity score equals 0, the word will be categorized as neutral. If the polarity score is more than 1, then the word will be categorized as positive, otherwise the word will be categorized as negative. In this study, due to low volume of dataset, the range of positive and negative values were altered to either more than 0 (>0) or less than 0 (<0). Algorithm 1 shows the steps of calculating the word polarity. The input is taken from the list of words that had went through feature extraction and selection phase. The output of Algorithm 1 will be an input to Algorithm 2 whereas Algorithm 2 is used to classify the word polarity score into specified categories: positive, negative and neutral. The last phase is results phase that presents the outcome of the analysis via data visualization tools such as word cloud, graphs and recommendation text.

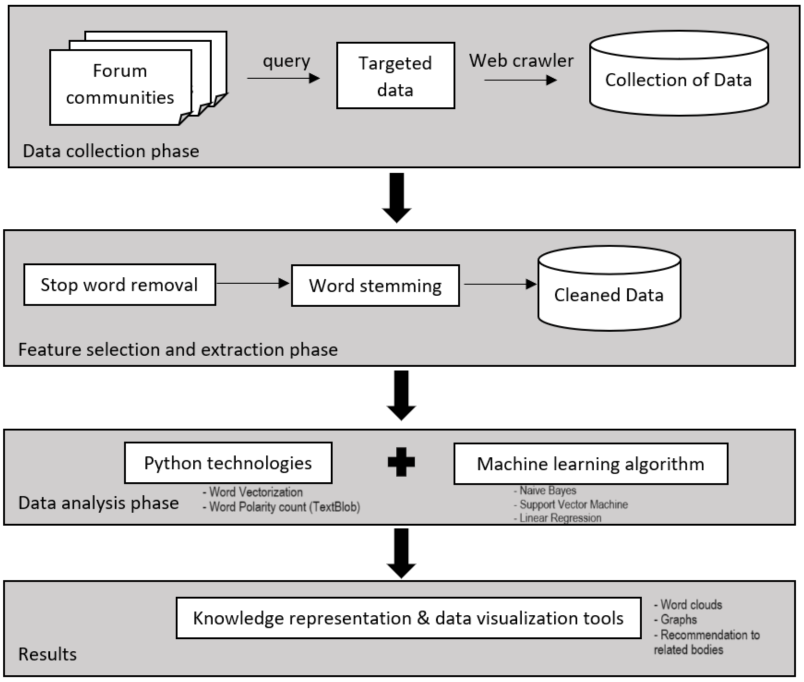

Figure 1. The proposed methodology on data analysis activities

Algorithm 1. Word polarity

```
Input: word (string): The input word for polarity analysis.
Output: score (float): Polarity score.
Procedure:
1. Function calculate_polarity_score(word):
   1.1. Parameters: word (string): The input word for polarity analysis.
   1.2. Calculate polarity_score using calculate_polarity_score(word).
   1.3. Returns: polarity_score (float): The polarity score of the word.
```

Algorithm 2. Word polarity classification

```
Input: score (float): The polarity word score for classification.
Output: category (string): category for each word.
Procedure:
2. Function categorize_word_polarity(word):
   2.1. Parameters: score (float): The polarity score for polarity analysis.
   2.2. If polarity_score equals 0:
        2.2.1. Set category to "neutral".
   2.3. Else, if polarity_score > 0:
        2.3.1. Set category to "positive".
   2.4. Else:
        2.4.1. Set category to "negative".
   2.5. Return category.
```

Instead of using word polarity in classifying the words, machine learning algorithms such as logistic regression (LR), NB, and SVM were also employed in classifying the words into positive, negative, and neutral categories. These algorithms have shown promising accuracy results in various prediction and classification problems [25]. Before the data went through the classification process, 75% of the data were split for training and 25% for testing dataset. These processes help to identify patterns within the collected data. Finally, the results of the analysis were visualized using data visualization tools such as graphs to illustrate the activities performed by data scientists during the COVID-19 pandemic. The common diagram that is used to visualize the word polarity is by using word cloud which can also be called from the Python library. The overall proposed process can be presented by the following mathematical modeling. The specific functions and decision variables can be further detailed based on the actual implementation and tools used in the sentiment analysis process. Let:

Q: set of specific queries used to gather the desired data.
D_{raw}: raw data collected from web forums.
$D_{processed}$: processed data after feature selection and extraction phases.
$D_{cleaned}$: data after stop word removal and word stemming processes.
V: set of words after word vectorization.
P: set of word polarities (positive, negative, neutral).
C: Set of word categories (positive, negative, neutral).

The entire process of data mining with sentiment analysis named as mining sentiment analysis that executes each process as a single function can be represented as shown in (1). The name of the function represented the process that will be conducted. For instance, WebScrapping(Q) represents the data scrapping process as mentioned in Q.

$$C = MiningSentimentAnalysis(Q) \tag{1}$$

Where:
D_{raw} = WebScraping(Q)
$D_{processed}$ = FeatureExtractSelect(D_{raw})
$D_{cleaned}$ = StoppingStemming($D_{processed}$)
V = WordVectorization($D_{cleaned}$)
P = TextBlobWordPolarity(V)
C = MachineLearningClassification(V)

3. RESULTS AND DISCUSSION

3.1. Data description

The data were collected from two web forums which are GitHub (https://github.com/) and Stack Overflow (https://stackoverflow.com/) from February 2020 until December 2021. 1,000 for GitHub and 3,685 for Stack Overflow of unstructured data have been collected. The keyword used in the scrapping process is “covid” for both websites to avoid bias. The dataset consists of unstructured type of data and being treated as text during the analysis process. The dataset has gone through the data analysis activities (feature selection, feature extraction, and sentiment analysis process) using Python technology. Instead of TextBlob, machine learning algorithms like NB, LR, and SVM were used to verify and improve the performance of the obtained results using TextBlob. Both data were split into 75:25 ratio for training and testing dataset.

3.2. Results discussions

The obtained results were visualized using visualization tools such as word cloud and table. Some recommendations were also discussed based on the findings. Table 1 lists the number of topics extracted from the GitHub and Stack Overflow using sentiment analysis approach by counting the sentiment score of each topic via TextBlob Python library. If the score is more than 0 (>0), the topic will be categorized as positive, meanwhile if the score equals to 0 (=0), the topic will be categorized as neutral, and if the score is less than 0 (<0), the topic will be categorized as negative. As can be seen, GitHub returned a very small size of topics under positive and negative categories. Not much can be extracted from both word clouds. Among the prominent words that appeared in positive category are “documentation”, “Laravel”, “data”, “python”, “rest-api” and “COVID19”. Meanwhile, in negative category, the most frequent words that occurred are “game”, “pygame”, “development”, “race”, and “covid”. In contrast, Stack Overflow page returned more topics compared to GitHub where 290 lies under positive topic category and 222 for negative topic category. The most prominent positive words for Stack Overflow are “using tableau”, “dashboard”, “covid record”, “detect covid” and “cases death”. Meanwhile for negative words, “plotly mapbox”, “mapbox base”, “base map”, “geojson file” and “ai covid” are among the frequent words that appear in this category.

Table 1. Number of words according to sentiment score categories

Web forum	GitHub	Stack Overflow
Neutral	993	3174
Positive	5	290
Negative	2	222

Tables 2 and 3 indicates the polarity scores of positive and negative topics for each GitHub and Stack Overflow datasets. For GitHub, most entries have positive sentiment scores, ranging from 0.033333 to 1 which cover a diverse range, including COVID-19 data analysis, gaming development, social aspects, and programming frameworks. The data appears to reflect a mix of technical and non-technical discussions, showcasing the variety of interests represented in the dataset. This may be due to the data scientists were involved with several activities such as discussions on collaboration between the communities, system, or application development, and sharing of information.

Table 2. Polarity scores of positive and negative categories for GitHub dataset

Topic	Polarity score	Category
[‘covid-race-game’, ‘game’, ‘pygame’, ‘deep-learning’, ‘artificial-intelligence’, ‘2d-game’, ‘python-game-development’, ‘covid-19’]	-0.4	Negative
[‘social’, ‘laravel’, ‘indonesia’, ‘indonesian’, ‘laravel-nova’, ‘covid-19’, ‘sekitarkita’]	0.033	Positive
[‘covid-19’, ‘coronavirus’, ‘api’, ‘react’, ‘documentation’, ‘node’, ‘rest-api’, ‘nextjs’, 'brazil', ‘now’, ‘zeit’, ‘free’, ‘brasil’, ‘dados’, ‘insomnia’, ‘dados-atualizados-sobre’]	0.4	Positive
[‘coronavirus’, ‘heroku’, ‘ncov’, ‘deaths’, ‘flask’, ‘python’, ‘api’, ‘confirmed’, 'recoveries', ‘covid-19’, ‘covid19-data’]	0.4	Positive
[‘covid-19’, ‘coronavirus’, ‘fitting’, ‘italy’, ‘epidemic’, ‘exponential-regression’, ‘sars-cov-2’]	0.5	Positive
[‘covid-19’, ‘coronavirus’, ‘covid19-data’, ‘awesome’, ‘corona’, ‘awesome-list’, ‘epidemiology’, ‘2019-ncov’, ‘2019ncov’, ‘coronavirus-info’, ‘covid19’, ‘sars-cov-2’, ‘awesome-corona’, ‘awesome-coronavirus’]	1	Positive

Table 3. Polarity scores of positive and negative categories for Stack Overflow dataset

Topic	Polarity score	Category
[‘plotly’, ‘mapbox’, ‘base’, ‘map’, ‘world’, ‘covid’, ‘cases’, ‘geojson’, ‘file’]	-0.8	Negative
[‘missing’, ‘authentication’, ‘token’, ‘c3’, ‘ai’, ‘covid’, ‘19’, ‘data’, ‘lake’]	-0.2	Negative
[‘importerror’, ‘cannot’, ‘imponame’, ‘covid’, ‘partially’, ‘initialized’, ‘module’, ‘covid’]	-0.1	Negative
[‘possible’, ‘use’, 'positive', ‘covid’, ‘record', 'prediction']	0.1136	Positive
[‘live’, ‘covid’, ‘data’, ‘dashboard’, ‘using’, ‘tableau’]	0.1363	Positive
[‘sound’, ‘features’, ‘detect’, ‘covid’, ‘19’]	0.4	Positive
[‘covid’, ‘return’, ‘latest’, ‘cases’, ‘deaths’, ‘data’, ‘without’, ‘filter’]	0.5	Positive

Meanwhile, Stack Overflow polarity scores involved a few discussions related to COVID-19, programming errors, and data visualization tools. The entries with negative sentiment scores (-0.8, -0.2, -0.1) suggest challenges or issues in programming related to COVID-19 data. Terms like ‘importerror’, ‘cannot’, and ‘partially initialized’ indicate potential difficulties in working with code or libraries. Entries with positive sentiment scores (0.1136364, 0.1363636, 0.4, 0.5) suggest a more favorable discussion, where topics include data visualization using tools like Tableau, sound features for detecting

COVID-19, and retrieving the latest cases and deaths data. The negative sentiment entries highlight specific technical challenges, such as authentication token issues, import errors, and module initialization problems. This suggests that developers or data analysts are struggling with obstacles in their work related to COVID-19 data processing and analysis.

Figures 2 and 3 present the word clouds for both GitHub and Stack Overflow website for all word categories. As can be seen from Figure 2, it can be concluded that instead of covid related words, there are a few words that depicted several activities that were conducted during the pandemic. The keywords were identified by looking at the words that represents verb. The keywords found were “corona virus tracking”, “tracker”, “dashboard”, “visualization”, and “hactoberfest”. Meanwhile, Figure 3 shows “covid tracker”, “plotting covid”, “data visualization”, “downloading covid”, and “android detect” keywords.

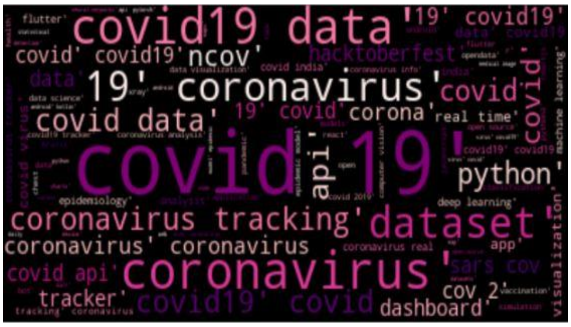

Figure 2. Word cloud from GitHub website

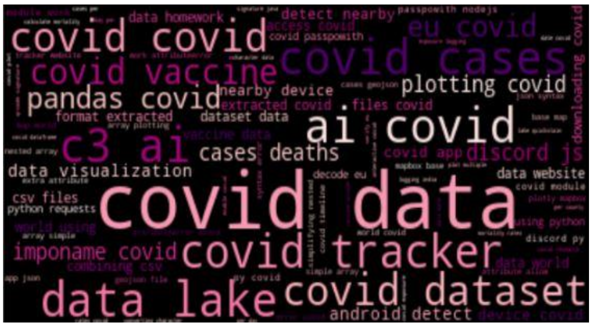

Figure 3. Word cloud from Stack Overflow website

3.3. Classification results

As mentioned in the previous subtopic, three machine learning models LR, NB, and SVM have been employed in the classification process. All models returned the same results which is 99.6% for GitHub dataset and 100% for Stack Overflow dataset with F1-score equals to 1. The obtained results indicate that the models achieved perfect precision and recall, with no false positives or false negatives. In conclusion, it seems that the three machine learning models (LR, NB, and SVM) are performing exceptionally well on both datasets. These models have effectively learned to classify the data with near-perfect accuracy. For future enhancement, it is essential to consider other factors such as the dataset size, data quality, and the potential for overfitting when interpreting these results in a real-world context.

4. CONCLUSION

This study introduces a robust research framework based on NLP and machine learning that analysis self-collected dataset on GitHub and Stack Overflow platforms about COVID-19 analytical projects engaged by data science community. Since recent literatures do not provide detail implementation framework for analyzing COVID-19 dataset based on NLP and machine learning, this framework that was constructed based the prepared COVID-19 dataset, provides a structured approach for extracting meaningful insights, sentiments, and patterns from textual data. The findings reveal that data scientists participated in numerous analytics related to the COVID-19 pandemic on GitHub and Stack Overflow, platforms that widely used by software developers and programmers. The study also revealed that a substantial number of data scientists dedicated their efforts to activities like tracking COVID-19 cases, creating data visualization dashboards, and developing COVID-19 tracking applications. Furthermore, evaluations of machine learning algorithms applied to sentiment classification within the GitHub and Stack Overflow datasets have shown a high performance that surpassed 90% accuracy rate as measured by the F1-score. In future, these algorithms can be used to demonstrate their potential for the development of automated response systems. The system would be capable of automatically addressing queries based on the sentiment expressed in any posts on platforms such as GitHub and Stack Overflow, or other platforms used by data scientists. This reporting approach has provided valuable insights to government agencies and decision-makers, aiding them in making effective choices. Due to this, government agencies could enhance their COVID-19 response by improving the communication strategies and allocating resources effectively in both technical and community engagement aspects.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Universitas Pembangunan Jaya (UPJ) for the funding of this project under the International Grant 2021 (Grant number: 100-TNCP/INT 17/6/2(054/2021).

REFERENCES

[1] K. Erol and T. Danyal, "Analysis of distance education activities conducted during COVID-19 pandemic," *Educational Research and Reviews*, vol. 15, no. 9, pp. 536–543, Sep. 2020, doi: 10.5897/ERR2020.4033.

[2] T. B. Hassen *et al.*, "Food behavior changes during the COVID-19 pandemic: statistical analysis of consumer survey data from Bosnia and Herzegovina," *Sustainability*, vol. 13, no. 15, p. 8617, Aug. 2021, doi: 10.3390/su13158617.

[3] K. Grondys, O. Ślusarczyk, H. I. Hussain, and A. Androniceanu, "Risk assessment of the SME sector operations during the COVID-19 pandemic," *International Journal of Environmental Research and Public Health*, vol. 18, no. 8, p. 4183, Apr. 2021, doi: 10.3390/ijerph18084183.

[4] N. Saxena, P. Gupta, R. Raman, and A. S. Rathore, "Role of data science in managing COVID-19 pandemic," *Indian Chemical Engineer*, vol. 62, no. 4, pp. 385–395, Oct. 2020, doi: 10.1080/00194506.2020.1855085.

[5] S. Latif *et al.*, "Leveraging data science to combat COVID-19: a comprehensive review," *IEEE Transactions on Artificial Intelligence*, vol. 1, no. 1, pp. 85–103, Aug. 2020, doi: 10.1109/TAI.2020.3020521.

[6] P. A. M. Oliveira, P. A. S. Neto, G. Silva, I. Ibiapina, W. L. Lira, and R. M. C. Andrade, "Software development during COVID-19 pandemic: an analysis of Stack Overflow and GitHub," in *2021 IEEE/ACM 3rd International Workshop on Software Engineering for Healthcare (SEH)*, Jun. 2021, pp. 5–12, doi: 10.1109/SEH52539.2021.00009.

[7] P. A. M. S. Neto *et al.*, "A deep dive into the impact of COVID-19 on software development," *IEEE Transactions on Software Engineering*, vol. 48, no. 9, pp. 3342–3360, Sep. 2022, doi: 10.1109/TSE.2021.3088759.

[8] V. Klotzman, F. Farmahinifarahani, and C. Lopes, "Public software development activity during the pandemic," in *Proceedings of the 15th ACM / IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, Oct. 2021, pp. 1–12, doi: 10.1145/3475716.3475778.

[9] E. Mbunge, B. Akinnuwesi, S. G. Fashoto, A. S. Metfula, and P. Mashwama, "A critical review of emerging technologies for tackling COVID-19 pandemic," *Human Behavior and Emerging Technologies*, vol. 3, no. 1, pp. 25–39, Jan. 2021, doi: 10.1002/hbe2.237.

[10] A. C. R. Tavares, N. A. Batista, and M. M. Moro, "How COVID-19 impacted data science: a topic retrieval and analysis from GitHub projects' descriptions," in *Anais do XXXVI Simpósio Brasileiro de Banco de Dados (SBBD 2021)*, Oct. 2021, pp. 325–330, doi: 10.5753/sbbd.2021.17893.

[11] K. Georgiou, N. Mittas, A. Chatzigeorgiou, and L. Angelis, "An empirical study of COVID-19 related posts on Stack Overflow: topics and technologies," *Journal of Systems and Software*, vol. 182, p. 111089, Dec. 2021, doi: 10.1016/j.jss.2021.111089.

[12] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: approaches, challenges and trends," *Knowledge-Based Systems*, vol. 226, p. 107134, Aug. 2021, doi: 10.1016/j.knosys.2021.107134.

[13] J.-N. Harba, G. Tigu, and A. A. Davidescu, "Exploring consumer emotions in pre-pandemic and pandemic times. a sentiment analysis of perceptions in the fine-dining restaurant industry in Bucharest, Romania," *International Journal of Environmental Research and Public Health*, vol. 18, no. 24, p. 13300, Dec. 2021, doi: 10.3390/ijerph182413300.

[14] R. Puertas, P. Carracedo, and L. Marti, "Environmental policies for the treatment of waste generated by COVID-19: text mining review," *Waste Management & Research: The Journal for a Sustainable Circular Economy*, vol. 40, no. 10, pp. 1480–1493, Oct. 2022, doi: 10.1177/0734242X221084073.

[15] P. Tansitpong, "Quality design for the COVID-19 pandemic: use of a web scraping technique on text comments and quality ratings from multiple online sources," in *International Series in Operations Research and Management Science*, vol. 320, Springer, 2022, pp. 329–341.

[16] L. Yang, Y. Li, J. Wang, and R. S. Sherratt, "Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning," *IEEE Access*, vol. 8, pp. 23522–23530, 2020, doi: 10.1109/ACCESS.2020.2969854.

[17] K. Kusum and S. P. Panda, "Sentiment analysis using global vector and long short-term memory," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 26, no. 1, pp. 414–422, Apr. 2022, doi: 10.11591/ijeecs.v26.i1.pp414-422.

[18] S. Vernikou, A. Lyras, and A. Kanavos, "Multiclass sentiment analysis on COVID-19-related tweets using deep learning models," *Neural Computing and Applications*, vol. 34, no. 22, pp. 19615–19627, Nov. 2022, doi: 10.1007/s00521-022-07650-2.

[19] D. Dangi, D. K. Dixit, and A. Bhagat, "Sentiment analysis of COVID-19 social media data through machine learning," *Multimedia Tools and Applications*, vol. 81, no. 29, pp. 42261–42283, Dec. 2022, doi: 10.1007/s11042-022-13492-w.

[20] X. Xiahou and Y. Harada, "B2C E-commerce customer churn prediction based on K-means and SVM," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 17, no. 2, pp. 458–475, Apr. 2022, doi: 10.3390/jtaer17020024.

[21] N. Z. Salih and W. Khalaf, "Prediction of student's performance through educational data mining techniques," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 3, pp. 1708–1715, Jun. 2021, doi: 10.11591/ijeecs.v22.i3.pp1708-1715.

[22] R. Andika and S. Suharjito, "Effects of using wordnet and spelling checker on classification methods in sentiment analysis for datasets using Bahasa," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 25, no. 3, pp. 1662–1671, Mar. 2022, doi: 10.11591/ijeecs.v25.i3.pp1662-1671.

[23] A. Alsaeedi and M. Zubair, "A study on sentiment analysis techniques of Twitter data," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 2, pp. 361–374, 2019, doi: 10.14569/IJACSA.2019.0100248.

[24] M. Khder, "Web scraping or web crawling: state of art, techniques, approaches and application," *International Journal of Advances in Soft Computing and its Applications*, vol. 13, no. 3, pp. 145–168, Dec. 2021, doi: 10.15849/IJASCA.211128.11.

[25] A. Chamekh, M. Mahfoudh, and G. Forestier, "Sentiment analysis based on deep learning in E-commerce," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13369 LNAI, 2022, pp. 498–507.

[26] A. K. Singh and M. Shashi, "Vectorization of text documents for identifying unifiable news articles," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 7, pp. 305–310, 2019, doi: 10.14569/IJACSA.2019.0100742.

BIOGRAPHIES OF AUTHORS

Masurah Mohamad is currently a senior lecturer at Universiti Teknologi MARA Perak Branch, Tapah Campus (UiTM Perak). She has received a B.Sc. (Hons.) in Computer Science majoring in Software Engineering from Universiti Teknologi Malaysia (UTM) in 2004, and M.Sc. in Computer Science also from Universiti Teknologi Malaysia (UTM) in 2006, respectively. She receives her Ph.D. in Computer Science field also from UTM in May 2021. She can be contacted at email: masur480@uitm.edu.my.

Dr. Suraya Masrom is an Associate Professor in Computing Science Studies at the College of Computing, Informatics, and Mathematics, University Teknologi MARA, Perak. She earned her Ph.D. in Information Technology and Quantitative Science from Universiti Teknologi MARA (UiTM) in 2015. Her career in the industry began as an Associate Network Engineer at Ramgate Systems Sdn. Bhd. (a subsidiary of DRB-HICOM) in June 1996 after completing her bachelor’s degree in computer science from Universiti Teknologi Malaysia (UTM). She can be contacted at email: suray078@uitm.edu.my.

Khairulliza Ahmad Salleh is a senior lecturer at the College of Computing, Informatics and Mathematics, Universiti Teknologi MARA (UiTM), Perak Branch, Tapah Campus, Malaysia. She receives her Ph.D. in Information Systems from The University of Auckland, New Zealand. She now serves as the Head of Centre of Studies. Her research interests include technology adoption, information security management and IT management. She can be contacted at email: khair279@uitm.edu.my.

Lathifah Alfat is a lecturer in Department of Informatics at Universitas Pembangunan Jaya, South Tangerang, Banten. She has received B.Eng. in Electrical Engineering majoring in Information Technology from Diponegoro University, Indonesia in 2015. She receives her Master of Engineering in Electrical Engineering majoring in Computer Engineering from University of Indonesia in 2020. She can be contacted at email: lathifah.alfat@upj.ac.id.

Muhammad Nasucha is an assistant professor in Department of Informatics at Universitas Pembangunan Jaya, South Tangerang, Banten. He has received his Bachelor of Engineering (with Honors) in Electrical Engineering majoring in Electronics and Telecommunications from Universitas Gadjah Mada in 1995. He received his M.Sc. in Electrical Engineering from Universitaet Kassel, Germany in 2004 and Ph.D. in Computation in Remote Sensing from Chiba University, Japan in 2020. He can be contacted at email: mohammad.nasucha@upj.ac.id.

Nur Uddin is an associate professor in Department of Informatics at Universitas Pembangunan Jaya, South Tangerang, Banten. Nur Uddin has received his Bachelor of Engineering in Aeronautics Engineering from Institut Teknologi Bandung at 2002 and M.Eng. in Mechanical Engineering, majoring in Control System from Gyeongsang National University, South Korea at 2009. In 2016, he receives his Ph.D. in Engineering Cybernetics from Norwegian University of Science and Technology, Norway. He can be contacted at email: nur.uddin55571@gmail.com.