Journal of Computer Science & Computational Mathematics, *Volume 13, Issue 1, March 2023*
DOI: 10.20967/jcscm.2023.01.003

Exploratory Data Analysis: Food Security Risk Among Twitter Users

Nur Azrawina Ahmad Kontar[1], Sofianita Mutalib[1,2,*], Haslizatul Fairuz Muhamed Hanum[1], Shuzlina Abdul-Rahman[1,2]

[1]School of Computing Sciences, College of Computing, Informatics and Media, Universiti Teknologi MARA,40450 Shah Alam, Selangor, Malaysia
[2]Research Initiative Group Intelligent Systems, Universiti Teknologi MARA,40450 Shah Alam, Selangor, Malaysia
[*]Corresponding author email: sofianita@uitm.edu.my

***Abstract*:** Food security poses a significant challenge in Malaysia, greatly impacting the well-being of its population. The objective of this study is to conduct a preliminary investigation into the application of sentiment analysis for predicting food security risks among Twitter users. By harnessing machine learning and natural language processing techniques, we aim to explore the emotions expressed in food-related tweets, gaining insights into prevailing sentiments and patterns within the Malaysian context. The study employed a descriptive method to analyze the tweets, compiling a comprehensive dataset of tweets in both Malay and English from Malaysian users. The tweets were then subjected to a cleaning process, which involved removing unusual characters, redundant tweets and converting the text to lower case. Through an exploratory analysis, we identified prevalent themes that could prove valuable for sentiment distribution within the dataset. Word cloud and histogram graphs were employed to identify the most common words related to food security. The sentiment scores of the tweets were also determined using Textblob. The preliminary findings indicate that sentiment analysis shows promise in identifying individuals at risk of food security issues. This research contributes to a nuanced understanding of food security challenges in Malaysia, providing valuable insights for policymakers and stakeholders to design targeted interventions. Further analysis and model refinement are planned to enhance the framework's effectiveness in detecting food insecurity risks among Malaysian Twitter users.

***Keywords*:** Descriptive analysis, food security, sentiment analysis, Twitter.

1. Introduction

Food security has been a persistent issue, particularly in developing nations, where food insecurity remains a significant concern. Food insecurity arises when individuals lacks consistent access to nutritious, safe, and sufficient food to support and maintain a healthy lifestyle [1]. According to a United Nations (UN) report in 2022, the total number of hungry people on the planet rose to 828 million in 2021, an increase of approximately 46 million since 2020 and 150 million more since the COVID-19 pandemic outbreak [2]. The factors contributing to food insecurity are complex. Some factors include poverty, increasing population, extreme climate changes, unemployment or low income, and conflict. Depending on access to resources that ensure the population's survival and availability, poverty and food insecurity (FI) affect people in every region worldwide to varying degrees [3].

Since food is a basic need, the demand for food should be less susceptible to crises than the demand for other goods and services. Nonetheless, the opposite of this happened when several researchers revealed that skyrocketing food prices occur due to several factors. Such factors include the disruption in food supply chains [4], export restrictions [5], and increased consumer demand during the global food crisis [6], especially in food-importing-dependent countries like Malaysia. In recent years, the rise of social media platforms, particularly Twitter, has served as a valuable forum for people to voice their ideas, concerns, and experiences about a variety of elements of their lives, including food-related issues. Twitter users frequently communicate their food-related experiences, ideas, and feelings, making it a potentially useful data source for academics looking to identify the risk of food insecurity.

Exploring Twitter conversations around food insecurity provides timely insight into important public health and social issues that are being discussed at an unprecedented moment [7]. Even though food poverty has consistently been a global issue, little attention has been paid to its monitoring with this technique [8]. This vast amount of data can provide policymakers with valuable information, especially in understanding human behaviour and emotions during such a crisis. Thus, sentiment analysis is widely recognized as a research area that extracts opinions, attitudes, and feelings from social media platforms like Twitter [9] and [10].

However, extracting food security events from social media can be challenging, partly because tweets are quite noisy and frequently lack enough context to identify food insecurity issues [11]. A large volume of raw data tends to be noisy due to relevant information, including repetitive content, hashtags, and slang. It may be difficult to recognise food security incidents because this useless information obscures the users' genuine sentiment. It may be difficult to recognise food security incidents because this useless information obscures the users' genuine sentiment. Plus, ambiguous wording in describing food insecurity concepts such as "hunger", "starvation," and "food shortage" has also added to the root problem. As a result, it may become challenging to pinpoint food insecurity based on user sentiment. Further elaborated by [12], food insecurity detection and prediction can be difficult due to the underlying network of variables ranging from food chain elements such as food availability, access, utilisation, and stability.

This paper presents the initial phase of a comprehensive study focusing on sentiment analysis for detecting food insecurity risk among Twitter users in Malaysia. To provide a solid foundation for subsequent analysis and to get insights into the emotional climate surrounding conversations about food security on Twitter, this paper focuses solely on data collection and early data exploration. The research objectives are as follows: (1) to identify the keywords for analyzing people's sentiments regarding food insecurity; and (2) to visualize food-insecurity-related keywords using word clouds and word frequency distribution.

This paper is structured as follows: Section 2 presents related works; Section 3 describes the methodology used in the experiment. Section 4 presents the findings and discussion of the experiment, and finally, the paper concludes with Section 5.

2. Related Works

Traditional methods such as surveys and questionnaires are considered tedious [13] in analysing the current trends of an issue within a country, and social media can be seen as a suitable alternative to conducting data acquisition. Several studies have been published on food insecurity's widespread dual cause and effect. Past researchers [14–17] have used household survey data from the statistics department to predict food security status. These researchers used a similar approach, using factors and characteristics of food security from the available data. Hence, this approach can be applied to detecting food insecurity risk using text classification based on keywords related to food insecurity from tweets.

One of the reasons why Twitter has become a favourite social platform for sentiment analysis and detecting such issues is that it serves as an excellent source of data for understanding the emotions of people from different societies. Millions of tweets are sent daily about practically any topic, contributing to the large volumes of user-generated content. Further elaborated by [18], it is an excellent source of public opinion and knowledge about the entire disaster lifecycle, including preparedness, response, mitigation, and resilience. Furthermore, Twitter offers the ability to conduct keyword searches that return tweets and every other relevant piece of information related to the phrase [19].

In [18], a group of researchers proposed a Twitter query string to effectively retrieve relevant tweets and assess topics within conversations, addressing how limited research in monitoring food security changes is based on public discourse on Twitter. They sought to improve upon previous work on food security topic modelling that used too broad or too narrow searches to retrieve all relevant food security tweets.

Another study by [8] used potential word pairings related to food insecurity in Uganda, such as "food Uganda," "hunger Uganda," "famine Uganda," and "drought Uganda." The primary challenge when conducting this study is related to data issues, specifically low tweet volume. As Uganda has a low number of Twitter users, this study took an alternative approach by using a worldwide dataset for opinion mining on food insecurity issues based on a Uganda case study. All tweets about food insecurity were packaged together and ready for trend development.

The research study by [20] utilised 138 important phrases that were manually selected to discuss food insecurity or insufficiency. Such terms include food availability, shortage, availability, acceptability, and sufficiency, which were then built into logical expressions. Due to emotional complexity, retrieving related tweets is insufficient for deducing people's emotions towards food insufficiency. Thus, a state-of-the-art NLP language model, Bidirectional Encoder Representations from Transformers (BERT), can identify the difference in sentiments and emotions portrayed in tweets by wholly understanding the context and meaning of words and phrases in a sentence.

To visualise the Tweets data, these researchers used WordCloud [21–23], a group of words represented in various sizes and colours based on their weight and significance, and word frequency distribution [24], which depicts how frequently each word appears in a text corpus. Unlike bar charts and pie charts, these two visualizations help readers better understand their results. The outcome of data exploratory study would be useful for further sentiment analysis with machine learning algorithms in wide area, as what had been done in literatures [25–27].

3. Methodology

This paper focuses on analyzing the sentiments of Malaysians towards the current food insecurity situation in the country. A collection of tweets from Malaysian users was gathered, and an exploratory data analysis was conducted to identify common topics, sentiment distribution, and significant trends.

3.1 Data Collection

A total of 81,458 tweets were collected in both Malay and English, starting on 1st February 2022 and ending on 31st March 2023. A list of keywords analyzed from previous research serves as a guideline to gather information about food insecurity to achieve this study's research objectives. Selected keywords related to topics of food insecurity, such as (1) food insufficiency, (2) unaffordability, (3) difficulty in finding food, and (4) food insecurity, were used as guidelines for search terms. For English tweets, the data has been scraped using 38 different search terms like “hunger,” “food insecure,” “food assistance,” “high food price,” and “skip meals” to increase the relevancy of the collected data. As for Malay tweets, such search terms were used including “bantuan makanan”, “kelaparan”, “lapar”, “ikat perut”, “harga barang naik”, “harga makanan naik”, “kekurangan bekalan makanan” and “makanan mahal”. Including both languages assured inclusivity and the representation of varied viewpoints from the Malaysian population.

Two different search approaches, based on location name and coordinates, were utilized to retrieve relevant tweets from specific states as well as the entire country of Malaysia. The combination of location-based and coordinate-based search methods enabled us to acquire tweets geotagged to particular states as well as those coming from Malaysia as a whole, offering a full picture of food-related interactions. For this purpose, fourteen states around Malaysia were chosen as one of the requirements besides the different search terms mentioned. Some of the features collected by *Snscrape* are datetime, tweet_id, username, text, and location. These features are predefined before the scraping process, as these are relevant to this study.

3.2 Data Preparation

The raw dataset consists of irrelevant elements and noise. During preprocessing by Python code, these elements are eliminated and replaced. This included removing URLs, unusual characters, numerals, emojis, mentions, and hashtags irrelevant to sentiment analysis. The process also included removing duplicate tweets and applying lowercase to text. Since tweet scraping is done in English and Malay, most Malay tweets are mixed with English. Thus, each tweet was translated into English to avoid inaccuracy during the sentiment analysis. These tweets will also be filtered for non-English tweets found in the tweet's dataset.

Following noise removal, we tokenized the remaining tweets, which involved breaking them down into individual words or tokens. Tokenization was a critical step in preparing the data for further analysis, allowing us to facilitate further analysis and ensure consistency in presenting textual data. Furthermore, stemming and lemmatization techniques were used to reduce words to their most basic forms, encouraging consistency in sentiment analysis. By reducing words to their base forms, we hoped to reflect the text's underlying meaning and emotional tone more precisely.

3.3 Visualization for Data Exploratory

The final stage involves presenting the processed text in a more meaningful representation. To achieve descriptive analytics outcome, two representations were selected: word cloud and word distribution, which is presented in the next section, providing visual insights and understanding of the prominent words and their distribution within the processed text.

4. Analysis and Discussion Results

4.1 Data Collection and Data Preparation

This section highlights the results of data collection and early data exploration for sentiment analysis of food security in Malaysia, based on identified keywords.

As can be seen in Table 1, despite using only eight search terms, there are more Malay tweets scraped than English tweets. Given that the majority of tweets are from Malaysians and Malay is the predominant language, this is presumably due to a demographic aspect. Table 2 summarises the overall number of tweets obtained via scraping using tools such as RapidMiner and Python (Snscrape).

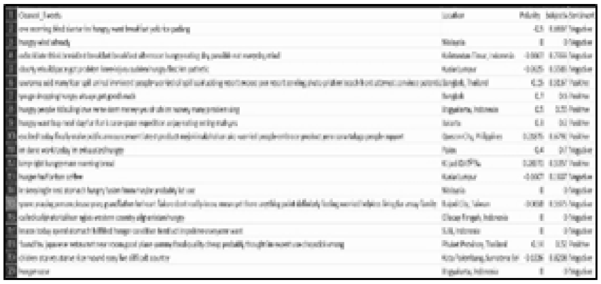
Table 1. Summary of tweets

Scraping tool	Location Search	Number of tweets
RapidMiner	Malaysia (Coordinates)	698
Python programming (Snscrape)	Malaysia (Location)	40,743
	Malaysia (Coordinates)	5,858
	States (Coordinates)	34,159
Total		**81,458**

Table 2. Comparison between raw and cleaned tweets by Malay and English

Tweets	***Before (Raw)***	***After (Cleaned)***
English	35,102	13,836
Malay	45,356	19,139
Total	**81,458**	**32,975**

A sample of tweets labelled with sentiments is shown in Fig. 1, there are 9,334 positive tweets, and the other 11,697 records are labelled as negative tweets. The score is extracted based on TextBlob method in Python.



Cleaned_Tweets	Location	Polarity	Subjectivity	Sentiment
[illegible]		[illegible]	[illegible]	Negative
[illegible]	Malaysia	0	0	Negative
[illegible]	[illegible]	[illegible]	[illegible]	Negative
[illegible]	Kuala Lumpur	[illegible]	[illegible]	Negative
[illegible]	Bangkok, Thailand	[illegible]	[illegible]	Positive
[illegible]	Bangkok	[illegible]	[illegible]	Positive
[illegible]	[illegible]	[illegible]	[illegible]	Positive
[illegible]	[illegible]	[illegible]	[illegible]	Positive
[illegible]	[illegible]	[illegible]	[illegible]	Positive
[illegible]	[illegible]	[illegible]	[illegible]	Negative
[illegible]	[illegible]	[illegible]	[illegible]	Positive
[illegible]	Kuala Lumpur	[illegible]	[illegible]	Negative
[illegible]	Malaysia	0	0	Negative
[illegible]	[illegible]	[illegible]	[illegible]	Negative
[illegible]	[illegible]	0	0	Negative
[illegible]	[illegible]	0	0	Negative
[illegible]	[illegible]	[illegible]	[illegible]	Positive
[illegible]	[illegible]	[illegible]	[illegible]	Negative
[illegible]	[illegible]	0	0	Negative

Figure 1. Sample Tweets Dataset from TextBlob output

The ID attribute was removed during text processing since it is not required in this analysis. This dataset combines all food security-related keywords based on a location-based search under Malaysia using Python programming *Snscrape*. Additional preprocessing steps include removing duplicate tweets and abbreviated words before combining different keyword datasets into one large dataset. This dataset is then used for early data exploration.

4.2 Data Exploratory

In this section, we report our preliminary findings from our data exploration phase to gain insights into the prevalent sentiments and themes regarding food insecurity among Malaysian Twitter users. To analyze the collected data, we employed various Python tools, including word cloud visualization and frequency distribution. These tools enable us to uncover meaningful patterns and trends related to food insecurity within the dataset.

As illustrated in Fig. 2, the word cloud visualizes the major themes and topics that emerged in discussions regarding food insecurity. Words like “hungry,” “worried,” “food,” and “price” highlight the prominence of conversations among Malaysian Twitter users regarding the challenges and issues related to food insecurity. These words signify the recurring themes and concerns expressed by users, emphasizing the significance of addressing food insecurity in Malaysia.

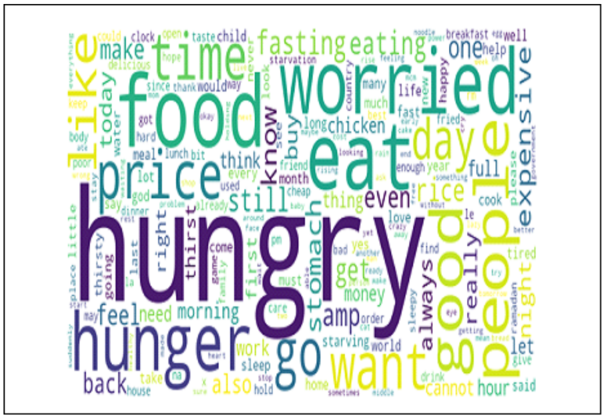


Figure 2. WordCloud of tweets collection related to food security

In addition to the word cloud, we performed a word frequency distribution analysis to quantify the frequency of certain words in the dataset, as depicted in Fig. 3. We determined the most frequently used keywords and expressions in debates about food security by ranking the words according to their frequency. The top 20 frequently used keywords found in the tweet’s dataset are displayed in this analysis. Obviously, word ‘hungry’ is the most dominant word in the word cloud and in the distribution graph. Based on Fig. 2, other words can be viewed in green colour, are ‘food’, ‘worried’, ‘eat’, ‘hunger’ and ‘price’. These words also appear among top frequent word in the distribution graph.

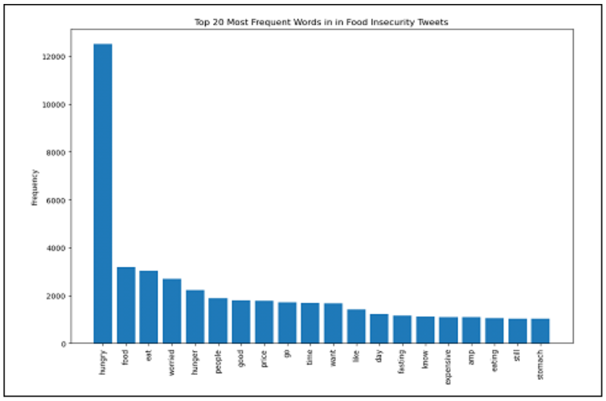


Figure 3. Word frequency distribution of tweets

This early exploratory analysis using word clouds and word frequency distribution gave us important information about the dataset’s common themes, key terms, and feelings. They act as a foundation for more in-depth analysis and sentiment classification algorithms in future.

5. Conclusion

This paper focuses on the preparation of the dataset for applying sentiment analysis to identify food insecurity on social media platforms. In the data collection phase, we collected a sample dataset of tweets from the Malaysian population, allowing us to capture localized viewpoints and experiences related to food insecurity. However, it is important to note that this paper is limited to the data acquisition and early data exploration phases. The preliminary findings highlight the importance of understanding and addressing food insecurity in Malaysia, as well as identifying individuals at risk of food insecurity and apprehend their emotional experiences. These findings provide valuable information for stakeholders and policymakers to guide further interventions. However, it is necessary for future research to refine the sentiment analysis framework and undertake more extensive analysis to evaluate its effectiveness. Additionally, incorporating contextual elements and socioeconomic indicators in future research can enhance the accuracy of food insecurity risk detection.

Acknowledgement

The authors would like to express the gratitude to Ministry of Higher Education, Malaysia for the research fund (FRGS/1/2021/ICT02/UITM/02/8) and School of Computing Sciences, College of Computing, Informatics and Media, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia for the research support.

References

[1] E. C. Lewis, U. Colón-Ramos, J. Gittelsohn, and L. Clay, “Food-seeking behaviors and food insecurity risk during the coronavirus disease 2019 pandemic,” *Journal of Nutrition Education and Behavior*, vol. 54, no. 2, pp. 159–171, 2022.

[2] Unicef *et al.*, “The state of food security and nutrition in the world (sofi) report-2022,” 2022.

[3] M. Pereira and A. M. Oliveira, “Poverty and food insecurity may increase as the threat of covid-19 spreads,” *Public health nutrition*, vol. 23, no. 17, pp. 3236–3240, 2020.

[4] D. Saccone, “Can the covid19 pandemic affect the achievement of the ‘zero hunger’goal? some preliminary reflections,” *The European Journal of Health Economics*, vol. 22, pp. 1025–1038, 2021.

[5] A. Panghal, R. S. Mor, S. S. Kamble, S. A. R. Khan, D. Kumar, and G. Soni, “Global food security post covid-19: Dearth or dwell in the developing world?” *Agronomy journal*, vol. 114, no. 1, pp. 878–884, 2022.

[6] S. Aday and M. S. Aday, “Impact of covid-19 on the food supply chain,” *Food Quality and Safety*, vol. 4, no. 4, pp. 167–180, 2020.

[7] F. Eskandari, A. A. Lake, and M. Butler, “Covid-19 pandemic and food poverty conversations: Social network analysis of twitter data,” *Nutrition Bulletin*, vol. 47, no. 1, pp. 93–105, 2022.

[8] A. Lukyamuzi, J. Ngubiri, and W. Okori, “Tracking food insecurity from tweets using data mining techniques,” in *Proceedings of the 2018 International Con-*

ference on Software Engineering in Africa, 2018, pp. 27–34.

[9] N. Yadav, O. Kudale, A. Rao, S. Gupta, and A. Shitole, "Twitter sentiment analysis using supervised machine learning," in *Intelligent data communication technologies and internet of things: Proceedings of ICICI 2020*. Springer, 2021, pp. 631–642.

[10] N. Yeasmin, N. I. Mahbub, M. K. Baowaly, B. C. Singh, Z. Alom, Z. Aung, and M. A. Azim, "Analysis and prediction of user sentiment on covid-19 pandemic using tweets," *Big Data and Cognitive Computing*, vol. 6, no. 2, p. 65, 2022.

[11] W. Gao, Y. Fang, Y. Wang, and F. Zhang, "Hrce: Detecting food security events in social media," in *Journal of Physics: Conference Series*, vol. 1437, no. 1. IOP Publishing, 2020, p. 012090.

[12] N. M. Martin, J. Sedoc, L. Poirier, A. J. Rosenblum, M. M. Reznar, J. Gittelsohn, and D. J. Barnett, "Harnessing artificial intelligence to improve food assistance: A scoping review of machine learning tools," 2022.

[13] N. Ahmad, Z. Alam, S. SK, and M. Husain, "Food insecurity: Concept, causes, effects and possible solutions," *IAR Journal of Humanities and Social Science*, vol. 2, no. 1, pp. 105–113, 2021.

[14] E. Ramadhani, B. Sartono, A. Hadi, T. Akhdansyah *et al.*, "Comparison of main characteristics of food insecurity using classification tree and random forest," *Sinkron: jurnal dan penelitian teknik informatika*, vol. 7, no. 4, pp. 2486–2497, 2022.

[15] A. Razzaq, U. I. Ahmed, S. Hashim, A. Hussain, S. Qadri, S. Ullah, A. Nawaz Shah, A. Imran, and A. Asghar, "An automatic determining food security status: Machine learning based analysis of household survey data," *International Journal of Food Properties*, vol. 24, no. 1, pp. 726–736, 2021.

[16] M. Alelign, T. M. Abuhay, A. Letta, and T. Dereje, "Identifying risk factors and predicting food security status using supervised machine learning techniques," in *2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*. IEEE, 2021, pp. 12–17.

[17] M. Nigus and H. Shashirekha, "A comparison of machine learning and deep learning models for predicting household food security status," *IJEER*, vol. 10, no. 2, pp. 308–311, 2022.

[18] N. M. Martin, L. Poirier, A. J. Rosenblum, M. M. Reznar, J. Gittelsohn, and D. J. Barnett, "Enhancing artificial intelligence for twitter-based public discourse on food security during the covid-19 pandemic," *Disaster medicine and public health preparedness*, pp. 1–25, 2022.

[19] R. Gupta, N. Gowalker, S. Joshi, and S. Patil, "Predicting risk in sentiment analysis using machine learning," *International Journal of Engineering and Advanced Technology*, vol. 9, no. 1, pp. 455–460, 2019.

[20] S. J. Goetz, C. Heaton, M. Imran, Y. Pan, Z. Tian, C. Schmidt, U. Qazi, F. Ofli, and P. Mitra, "Food insufficiency and twitter emotions during a pandemic," *Applied Economic Perspectives and Policy*, vol. 45, no. 2, pp. 1189–1210, 2023.

[21] R. A. Sabaruddin and S. Saee, "Malay tweets: discovering mental health situation during covid-19 pandemic in malaysia," in *2021 IEEE 19th Student Conference on Research and Development (SCOReD)*. IEEE, 2021, pp. 58–63.

[22] J. Samuel, G. M. N. Ali, M. M. Rahman, E. Esawi, and Y. Samuel, "Covid-19 public sentiment insights and machine learning for tweets classification," *Information*, vol. 11, no. 6, p. 314, 2020.

[23] H. Yin, S. Yang, and J. Li, "Detecting topic and sentiment dynamics due to covid-19 pandemic using social media," in *Advanced Data Mining and Applications: 16th International Conference, ADMA 2020, Foshan, China, November 12–14, 2020, Proceedings 16*. Springer, 2020, pp. 610–623.

[24] T. Rahman and S. Aktar, "A machine learning approach to track covid-19 pandemic using sentiment analysis," in *2021 3rd International Conference on Electrical & Electronic Engineering (ICEEE)*. IEEE, 2021, pp. 145–148.

[25] A. Nabiha, S. Mutalib, and A. M. Ab Malik, "Sentiment analysis for informal malay text in social commerce," in *2021 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*. IEEE, 2021, pp. 1–6.

[26] N. A. S. Remali, M. R. Shamsuddin, and S. Abdul-Rahman, "Sentiment analysis on online learning for higher education during covid-19," in *2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*. IEEE, 2022, pp. 142–147.

[27] M. R. A. Rahim, S. Abdul-Rahman, and Y. Mahmud, "Customers' opinions on mobile telecommunication services in malaysia using sentiment analysis," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 12, 2021.