Implementation of Text Processing Techniques on Citizen Opinions Regarding Floods in Surabaya

Rony Kriswibowo[1], Putri Ariatna Alia[2], Johan Suryo Prayogo[3], Rusina Widha Febriana[4]

[12,3,4] Universitas Anwar Medika; email: rony.kriswibowo@uam.ac.id, putriariatna@gmail.com, jodimasjolie@gmail.com , widha.ennvil@gmail.com

[Received: 28 February 2024, Revised: 13 May 2024, Accepted: 27 May 2024]
Corresponding Author: Rony Kriswibowo

ABSTRACT — An analysis of residents opinions on flooding in Surabaya is essential to identify their perceptions and aspirations towards this disaster. This can facilitate the making of appropriate flood mitigation policies. The current problem is that it is difficult to retrieve data from X in the form of both users and tweets automatically. So that in some studies that use tweet data becomes less efficient in the data collection process. This research confirmed the use of messages to determine highly impactful disaster zones and showed how tweets can be used to identify oscillations in disaster intensity over time. The topic of this research is to apply the Data Crawling method to obtain datasets on social media X. Then the next method is text preprocessing using Wordcloud, Matplotlib, (NTLK) Natural Language Toolkit, and Sastrawi libraries. In natural language processing, the data to be extracted includes unstructured or “arbitrary” data. In normal dialect preparing (NLP), the information to be extracted includes unstructured or "self-assertive" information. For future purposes (assumption examination, subject modeling, etc.), such information must be changed over into organized data. The discoveries of the think about can help the organization in comprehending the necessities and inclinations of the people with respect to surges. This article demonstrates how Artificial Intelegence may be applied to text data analysis in order to provide insightful findings. the outcomes of this research can help the government in making more effective policies to overcome flooding in Surabaya.

KEYWORD — X, Crawling Data, Text Preprocessing, Natural language

I. INTRODUCTION

East Java regions Sidoarjo and Surabaya are seeing significant population growth and high construction densities. According to information gathered from the field, the Waru highway serves as the primary route for travelers traveling from Sidoarjo to Surabaya [1]. Furthermore, the location serves as one of the routes leading to Purabaya Terminal. This makes it difficult for the community to go to events, which leads to traffic congestion in the Sidoarjo area.

A flood is a type of natural catastrophe that arises when an excessive amount of water overflows land, typically due to factors such as elevated rainfall, excessive groundwater extraction, clogged sewers, and other causes [2]. In addition to unfilled drains, another source of flooding is an elevation in sea level higher than the land. [3]. Social media is widely used by individuals to exchange information and voice opinions on the flooding tragedy that occurred in an East Javan district. X is one social media platform that's used [4].

X users' responses to the East Java flood catastrophe were varied. From the start of the year to February 2024, a large number of tweets were sent on the floods that affected different parts of East Java. X users' most common opinions are those on how local flood catastrophes are handled [5]. One of the world's most active X users is found in Indonesia [6].

Finding out what the public`s sentiment was like following the flood is the aim of this investigation. One possible way to reflect citizen knowledge is through the use of X. This study confirms the use of geolocation messaging to demarcate highly impacting catastrophe zones and demonstrates how tweets may be used to identify oscillations in disaster intensity over time [7].

The topic in this research is to apply the Data Crawling method to get datasets on social media X. This is because twitter makes it easy to read, write and collect data containing temporal and spatial information [16]. Google Collaboratory is a web-based interactive IDE made by Google that runs in a cloud environment. It allows us to write and run code interactively through a browser [8]. Electronic devices such as cell phones, laptops, and notebooks are not only used to communicate, but also to obtain information about the surrounding situation. People get information and various social media [9].

This background informs the research technique, which uses text preprocessing to examine the sentiment analysis method of X data. When it comes to natural language processing (NLP), the data that has to be extracted includes unstructured or "arbitrary" data. For future purposes (sentiment analysis, topic modeling, etc.), it must thus be transformed into structured data.

II. LITERATURE REVIEW

The issue brought up was the flood calamity that struck West Java in January 2020, according to earlier study by Aalzaz et al (2021). This study examined the opinions of the general people on X on the management of flood catastrophes in West Java in January 2020. Classifying tweets that expressed public opinion on how flood catastrophes are handled allowed for the study to be done. The confusion matrix was employed as the validation technique in this study.[9].

Then research on flood sentiment analysis by Rahmat Hartawan et al (2023) The purpose of this study is to examine public opinion on the efficacy of Malang City`s Early Warning System (EWS) for flooding disasters. The original data will be weighted using TF-IDF for sentiment analysis. The efficacy of flood EWS in Malang City is still seen negatively by the general population, nonetheless [10]. Further research by Trida et al (2021) In his research, he found that floods received more attention on X than landslides. In the implementation of disaster management, we argue that X data can be used in all phases. X data in disaster management, especially hydrometeorological disasters, has weaknesses such as it can only be used in big cities in Java Island. Another weakness is that information from X data has not been confirmed whether it represents the same demographic characteristics as the conditions in the field and the level of validity of information that cannot be accounted for [11].

Further pertinent study by Mera et al (2021) uses machine learning techniques to classify tweets, particularly those concerning flood catastrophes. A number of data pretreatment procedures were used to begin the categorization of flood catastrophe communications, and they included feature extraction and word weighting from X data [12].

Study conducted by Finki and Mambang (2022) on Natural Language Processing (NLP). In order to raise awareness, track catastrophes, and expedite flood disaster recovery using social media analysis, the goal of this project is to develop a visualization model for flood response information from disaster management. Data sources produced from Instagram postings are utilized for analysis in the Natural Language Processing (NLP) technique. Utilizing Natural Language Processing (NLP) as a process step, data sources from Instagram with flood hastags in Kalimantan are leveraged to get the essential information visualization components for expediting flood response data. The process of visualizing social media data by extraction of information from Instagram posts, comments, and hashtags speeds up information retrieval for issue resolution and emphasizes the significance of flood reaction information velocity [13].

To obtain data in JSON from X, the crawler employs HTTP GET along with application credentials that are created from the X Dev Console. Following the Crawler`s receipt of results from X, the JSON is parsed into a local slice of a proprietary data type. Next, the Crawler goes over each component of that slice, processing it to eliminate tweets where the language wasn't recognized or the location wasn't supplied by the user [14]. Crawling data on X can use two search systems, by user and by keyword. Search using by keyword is a search using fragments of words or hashtags with a total of 100 tweets downloaded in one process. downloaded in one process is a maximum of 100 tweets. While searching by user, namely search based on the X user account name with a maximum of 200 tweets downloaded in one process maximum 200 tweets [15].

X is the biggest rival to Facebook in terms of active users. This is reflected in the large number of links to X included in the BlogIntelligence data collection. We prefer to scan X users because, based on our experience, we believe that the majority of Bloggers also keep a X account where they share their writings and other ideas. However, the next natural step would be to crawl these people' tweets as well as any other connected information. The next portion of this page describes the opportunities provided by X's APIs. After that, we discuss how the crawling procedure was put into practice [8].

The majority of machine learning-based sentiment analysis experiments seen in literature involve supervised learning and require sufficiently big pre-labeled datasets in certain domains. It is obvious that this activity is laborious, costly, and time-consuming to construct, and handling unknown data is challenging. With its small datasets, this work tackles semi-supervised learning for sentiment analysis in Vietnamese. We have compiled a list of several preprocessing methods that were used to clean and standardize data, handle negations, and handle intensification to enhance speed. Additionally, methods for data augmentation have been introduced, which create new data from the original data to improve training data automatically. We have conducted several tests and produced competitive outcomes that might serve as inspiration for future ideas [17]. Text pre-processing is the first stage in data processing so that data can be processed and data is ready for the research process, one of which is sentiment analysis. Text pre-processing can include cleaning, casefolding, tokenizing, stemming, and filtering.

The method of text processing has been shown to be very effective in obtaining sentiment analysis results from social media X based on relevant prior research. The drawback of previous research was that data collection was limited to 100 to 200 tweets. The novelty of this study is that researchers have analysed the phenomenon of flooding in Surabaya, Indonesia, which occurred in 2024.

The current generation of students is studying and doing research in the field of machine learning and neural networks, mostly due to the buzz surrounding these subjects. However, a major obstacle that many of these researchers and students confront is the lack of adequate infrastructure. A sufficient number and quality of CPUs and GPUs are needed to run complex algorithms. Fortunately, Google offers a free resource called Google Collaboratory to help with such requests. However, consumers' ability to supply the necessary data for computational purposes is limited. This study explores how researchers might use Google Collaboration's manual upload capability to enter and retrieve data dynamically. This will enable scholars or students to focus more intently on their argument rather than focusing on the data [18].

III. METHODOLOGY

The initial stage is to look for a literature study on relevant research. Next determined the method to be used based on previous research sources and the results of a literature review. Then data collection and data analysis produce conclusions. The application of the method is carried out at the 5th stage, namely data analysis using Text Processing.

The stages in this research use data crawling and text preprocessing methods, after the literature study then determining the method, the data crawling stage is at this stage, namely data collection. Then the next stage is data analysis. This analysis data uses the text preprocessing method. Then the conclusion is drawn. Flowchart for crawling data is shown in Figure 5 then the data analysis stage in Figure 6. The following picture of the research flow in the form of a flowchart is shown in Figure 1.

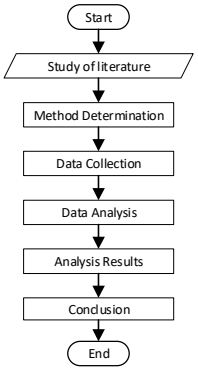

Figure 1. Methodology Research

The first stage of this research is crawling data on X [19]. This research uses the data crawling source code as follows, the first step is to retrieve the auth X token. The following is the source code:

```
#@title X Auth Token
X_auth_token = 'Auth Key X'
```

Figure 2. Source Code X Auth Token

Then after getting the X Auth Token, the next step is to install the python library, following the source code:

```
# Import required Python package
!pip install pandas

# Install Node.js (because tweet-harvest built using Node.js)
!sudo apt-get update
!sudo apt-get install -y ca-certificates curl gnupg
!sudo mkdir -p /etc/apt/keyrings
!curl -fsSL https://deb.nodesource.com/gpgkey/nodesource-repo.gpg.key |
sudo gpg --dearmor -o /etc/apt/keyrings/nodesource.gpg

!NODE_MAJOR=20 && echo "deb [signed-by=/etc/apt/keyrings/nodesource.gpg]
https://deb.nodesource.com/node_$NODE_MAJOR.x nodistro main" | sudo tee
/etc/apt/sources.list.d/nodesource.list

!sudo apt-get update
!sudo apt-get install nodejs -y
```

Figure 3. Source Code Import Library Python

Next is the process of crawling data based on X, here is the source code:

```
# Crawl Data

filename = 'Banjir.csv'
search_keyword = 'Banjir Sidoarjo lang:id'
limit = 1000

!npx --yes tweet-harvest@2.2.8 -o "{filename}" -s "{search_keyword}" -l
{limit} --token {X auth token}
```

Figure 4. Source Code Crawl Data

Here are the stages of data collection from X:

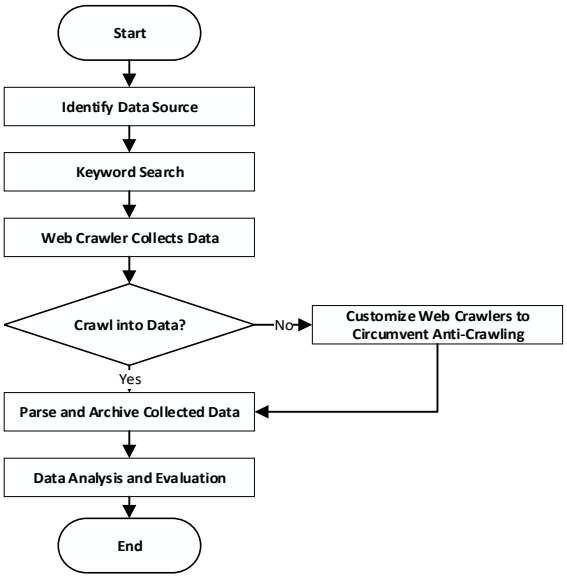

Figure 5. Flowchart Crawling data from X

After crawling the data, a dataset from X was obtained. This dataset is then text preprocessed to get a clearer visualization of the data [20].

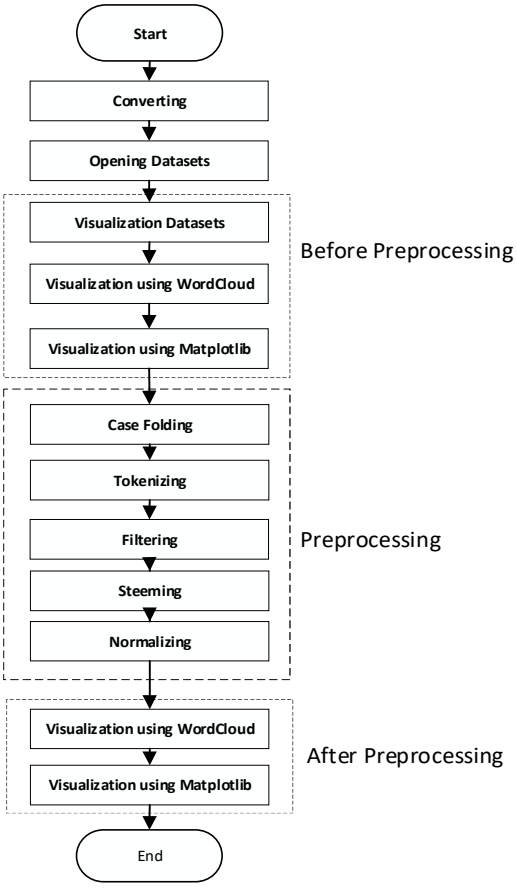

Figure 6. Flowchart Text Preprosessing

IV. RESULTS AND DISCUSSION

The first preprocessing stage is to get the crawling results file from X, then convert it into csv format. In the preprocessing process, there are stages before preprocessing and after preprocessing to get better and clearer data visualization. In the stage before preprocessing, dataset visualization is carried out, then raw data visualization is carried out before preprocessing. Natural language processing techniques such as keyword weighting, sentiment classification, and opinion clustering were able to map and categorize the opinions of Surabaya residents regarding their perceptions and experiences with flooding. The following is an image of the processing results before preprocessing displayed in the form of a wordcloud.

Figure 7. Data visualization before preprocessing using Wordcloud

The next step of the process is to display the datasets in graphical form.

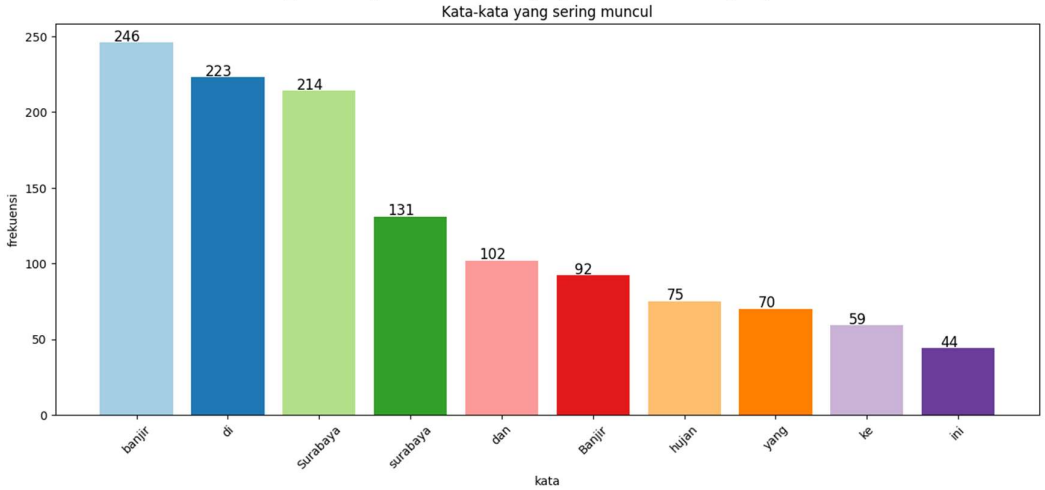

Figure 8. Data visualization before preprocessing using Matplotlib

The results visualized before preprocessing are the visualization of raw data getting the word that often appears is '**banjir**' with a total of 246 words, then followed by the word '**di**' with a total of 223 words '**Surabaya**' a total of 214 words '**surabaya**' a total of 131. So it can be concluded that the word mix is still the same word, so there is a need for text preprocessing to reduce the same word mix that often appears. The next step of the process is Text pre-processing can include cleaning, casefolding, tokenizing, stemming, and filtering. The current phenomenon of flooding in Surabaya in 2024 is a novelty of this research. In 2024, no other relevant research was carried out on a sentiment analysis of flood events in Surabaya.

Figure 9. Data visualization after preprocessing using Wordcloud

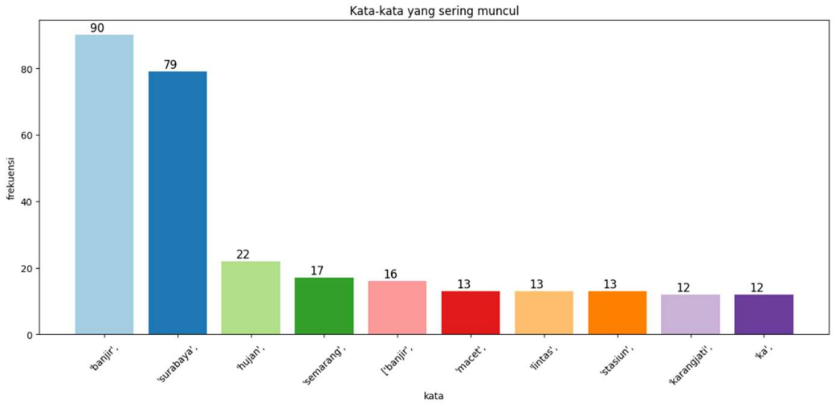

Figure 10. Data visualization after preprocessing using Matplotlib

The results visualized after preprocessing are the visualization of raw data getting the word that often appears is '**banjir**' with a total of 90 words, then followed by the word '**surabaya**' with a total of 79 words '**hujan**' a total of 22 words '**semarang**' a total of 17. Thus, preprocessing reduces the frequency of frequently occurring double words. Thus, it may be said that the term that comes up most frequently is flood. This indicates that people's opinions expressed on social media platforms like X frequently mention Banjir. There are 2 process stages in text processing, namely the first data results before the casefolding, tokenizing, filtering, steeming and normalizing methods are carried out. Here I show the comparison before and after visual processing.

Table 1 Result of Text Preprocessing

Word	Before Preprocessing	After Preprocessing
Banjir	246	90
di	223	0
hujan	75	22
surabaya	214	79
macet	0	13

So there are 2 stages before and after aiming to eliminate some problems that can interfere with data processing. This is because a lot of data has inconsistent formats. making the data accurate, consistent, and suitable for analysis..

V. CONCLUSION

Public opinion analysis can be done using text processing methods. By offering a platform for cloud computing, the use of Google Collaboratory eases the analysis process. The findings from this research can help the government in understanding the needs and preferences of the community regarding flooding. This article shows how AI can be applied to text data analysis to provide in-depth findings. Data analysis becomes more sophisticated and effective by using Google Collaboratory. The results of this research can help the government in making more effective policies to overcome flooding in Surabaya.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] Z. H. Yuyun Tia Triana, "Kajian potensi daerah rawan banjir rob dan adaptasi masyarakat di wilayah pesisir utara surabaya," *Juv. J. Ilm. Kelaut. dan Perikan.*, vol. 1, no. 1, pp. 141–150, 2020.
[2] S. Rizki Aditiya, "Kajian banjir di wilayah ketintang surabaya," *J. Rekayasa dan Manaj. Konstr.*, vol. 7, no. 2, pp. 157–162, 2019.
[3] A. A. Lynda Refnitasari, Hendra Wahyu Cahyaka, Krisna Dwi Handayani, "Analisis Kerentanan Fisik Wilayah Pesisir Utara Kota Surabaya Terhadap Bencana Banjir Rob," *J. Tata Kota dan Drh.*, vol. 14, no. 2, pp. 55–62, 2022.

[4] L. K. Katherina, "Dinamika Pertumbuhan Penduduk Dan Kejadian Banjir Di Kota : Kasus Surabaya (Dynamic Of Population Growth And Flooding Incidents In Cities : Case Of Surabaya)," *J. Kependud. Indones.*, vol. 12, no. 2, pp. 131–144, 2017.
[5] A. R. Abdillah, F. N. Hasan, T. Informatika, F. N. Hasan, D. Mining, and A. Sentimen, "Analisis Sentimen Terhadap Kandidat Calon Presiden Berdasarkan Tweets Di Sosial Media Menggunakan Naive Bayes Classifier Sentiment Analysis of Presidential Candidates Based on Tweets on," *SMATIKA J. STIKI Inform. J.*, vol. 13, no. 1, pp. 117–130, 2024.
[6] I. Rustanto, "Media Sentiment Analysis of East Java Province : Lexicon-Based vs Machine Learning," *IPTEK J. Proc. Ser.*, vol. 1, no. 1, pp. 203–208, 2019.
[7] N. T. Putri, I. D. Wijaya, A. Retno, and T. Hayati, "Analisis Sentimen Opini Masyarakat Terhadap Pembangunan Infrastruktur Kota Malang Melalui Twiter Dengan Menggunakan Metode Support Vector Machine," *2020 Semin. Inform. Apl. Polinema - 2020*, pp. 118–123, 2020.
[8] P. Berger, P. Hennig, T. Klingbeil, M. Kohnen, S. Pade, and C. Meinel, "Mining the Boundaries of Social Networks : Crawling Facebook and Twitter for BlogIntelligence," pp. 223–229.
[9] A. L. S. A.-Z. Gunawan, Jondri, and K. M. Lhaksamana, "Analisis Sentimen pada Media Sosial Twitter terhadap Penanganan Bencana Banjir di Jawa Barat dengan Metode Jaringan Saraf Tiruan Sentiment," *e-Proceeding Eng.*, vol. 8, no. 2, p. 2965, 2021, [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/viewFile/14695/14472
[10] S. K. A. Rahmat Hartawan, Muhammad Faza Imani Putra, "Analisis sentimen persepsi masyarakat terhadap efektivitas early warning system bencana banjir di kota malang menggunakan pembobotan tf - idf 1," *Pangripta J. Ilm. Kaji. Perenc. Pembang.*, vol. 6, no. 2, pp. 99–108, 2023.
[11] T. Ridho Fariz, S. Suhardono, and S. Verdiana, "Pemanfaatan Data Twitter Dalam Penanggulangan Bencana Banjir dan Longsor Use of Twitter Data in Flood and Landslide Disaster Management," *Cogito Smart J.* |, vol. 7, no. 1, pp. 135–147, 2021.
[12] M. K. Delimayanti, R. Sari, M. Laya, M. R. Faisal, and P. Pahrul, "Pemanfaatan Metode Multiclass-SVM pada Model Klasifikasi Pesan Bencana Banjir di Twitter," *Edu Komputika J.*, vol. 8, no. 1, pp. 39–47, 2021, doi: 10.15294/edukomputika.v8i1.47858.
[13] F. D. Marleny and Mambang, "Sosial Media Analisis Berbasis NLP Untuk Mempercepat Tanggap Bencana Banjir," *Tematik*, vol. 9, no. 1, pp. 1–7, 2022, doi: 10.38204/tematik.v9i1.897.
[14] M. A. A. Shahab Saquib Sohail, Mohammad Muzammil Khan, Mohd Arsalan, Aslam Khan, Jamshed Siddiqui, Syed Hamid Hasan, "Crawling Twitter data through API : A technical / legal perspective," *arXiv Prepr. arXiv2105.10724.*, pp. 1–8, 2021.
[15] J. E. Sembodo, E. B. Setiawan, and Z. K. A. Baizal, "Data Crawling Otomatis pada Twitter," *Ind. Symp. Comput.*, no. August, pp. 11–16, 2016, doi: 10.21108/indosc.2016.111.
[16] V. A. Shahputri and Y. Yamasari, "Analisis Sentimen Mengenai Pasca Bencana Alam Menggunakan Metode K-Nearest Neighbor (K-NN) dan Decision Tree," *J. Informatics Comput. Sci.*, vol. 05, pp. 377–388, 2023.
[17] H. T. Duong, T. Anh, and N. Thi, "A review : preprocessing techniques and data augmentation for sentiment analysis," *Comput. Soc. Networks*, pp. 1–16, 2021, doi: 10.1186/s40649-020-00080-x.
[18] S. V Halyal, "Running Google Colaboratory as a server – transferring dynamic data in and out of colabs," *I.J. Educ. Manag. Eng.*, no. September, pp. 35–39, 2019, doi: 10.5815/ijeme.2019.06.04.
[19] Z. Ma, "Research on Twitter Data Crawling and Data Visualization Analysis Based on Python," in *Computing and Data Science*, 2021, pp. 351–362.
[20] T. D. Dikiyanti, A. M. Rukmi, and M. I. Irawan, "Sentiment analysis and topic modeling of BPJS Kesehatan based on twitter crawling data using Indonesian Sentiment Lexicon and Latent Dirichlet Allocation algorithm," *J. Phys. Conf. Ser.*, vol. 1821, no. 1, 2021, doi: 10.1088/1742-6596/1821/1/012054.