CoreID JOURNAL

Vol. 1 No. 1, July 2023 (66-74)
ISSN: 2987-6990
DOI: https://doi.org/10.60005/coreid.v1i2.14

# Sentiment Analysis on Twitter Social Media Regarding Depression Disorder Using the Naive Bayes Method

**Nur Lickha Lavenia[1], Reisa Permatasari[2]**
[1,2]Information System, Computer Science Faculty, UPN Veteran Jawa Timur

**Article Info**

***Article history:***

***Keywords:***

Depression disorder
Twitter
Naïve Bayes

**ABSTRACT**

Depression disorder is a serious issue in mental health that affects many individuals worldwide. This research analyzes the sentiments related to depression disorder on Twitter using the Naïve Bayes method. Depression-related tweet data was collected through snscrape and processed to eliminate irrelevant information. Three Naïve Bayes methods, namely Multinomial, Gaussian, and Bernoulli, were compared to classify positive, negative, or neutral sentiments in each tweet. The results of the study indicate that Multinomial Naïve Bayes exhibited the best performance with an accuracy rate of 90.13%, followed by Gaussian Naïve Bayes (88.38%), and Bernoulli Naïve Bayes (85.37%).



***Corresponding Author:***

Nur Lickha Lavenia
Information System, UPN Veteran Jawa Timur,
Jl. Semampir Selatan IIA No.114, Medokan Semampir, Kec. Sukolilo, Surabaya, Jawa Timur 60119
Email: lickhalavenia13@gmail.com

## 1. INTRODUCTION

The internet has become an essential part of nearly everyone's daily life. Internet usage in Indonesia reached 202 million people, approximately 73.7% of the total population, in 2021. Popular social media platforms in Indonesia include WhatsApp (90 million monthly active users), Instagram (78 million monthly active users), and Facebook (73 million monthly active users). TikTok is also gaining popularity with 62 million monthly active users (Source: Central Statistics Agency, 2021; We Are Social and Hootsuite, 2022).

Twitter is popular in Indonesia, with 19 million monthly active users in 2022 (We Are Social and Hootsuite). Through Twitter, individuals can share personal experiences related to mental health, including depression disorder. Research indicates that Twitter can identify behavioral patterns and emotions related to depression disorder (Al-Qaysi and Al-Janabi, 2020). Indonesian Twitter users often discuss mental health topics, including depression (Santoso et al., 2020). Twitter can serve as a source of information and support for individuals with depression disorder in Indonesia.

Depression disorder is a serious mental health issue affecting millions of people, including those in Indonesia. In 2018, approximately 3.8 million people in Indonesia were diagnosed with depression disorder (Ministry of Health of the Republic of Indonesia, 2018). Depression disorder can lead to feelings of sadness, despair, lack of motivation, and disruption of daily life. Many individuals seek support on social media platforms, including Twitter. In sentiment analysis on Twitter related to depression disorder, researchers used the Naïve Bayes algorithm based on previous studies showing its superior accuracy (Widaningsih, 2019).

This research aims to understand public perspectives on depression disorder through sentiment analysis of relevant tweets. The results of sentiment analysis are expected to reveal positive, negative, or neutral views of the public towards depression disorder. Additionally, sentiment analysis helps identify issues related

to depression disorder, such as stigma, lack of support, or lack of knowledge. The findings of this research can assist mental health professionals and policymakers in understanding the perspectives and needs of the public regarding depression disorder and in designing more effective programs and interventions to increase awareness and understanding of depression disorder among the public

## 2. METHOD

The steps used in this research are outlined in the research methodology chapter, providing a clear research structure. In this chapter, the explanation is presented in the following research flow:


Figure 1. Research flow

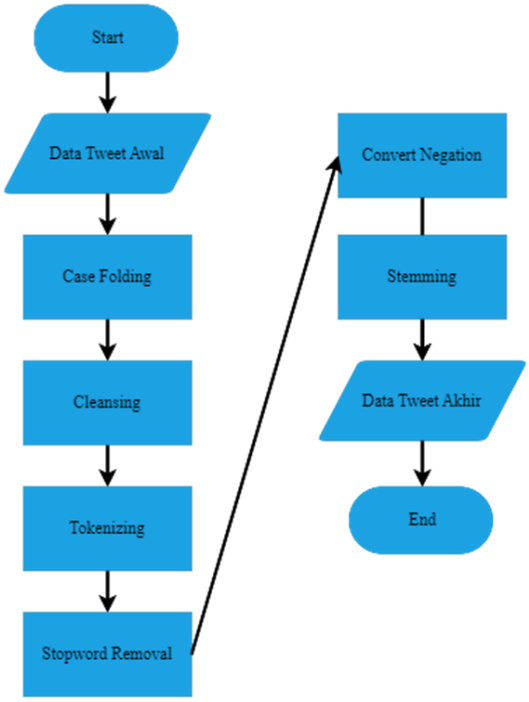


Figure 2. Classification Model Design

## 3. RESULTS AND DISCUSSION

This chapter discusses the results and findings of the research conducted on Twitter sentiment analysis regarding depression disorder using the Naive Bayes method. The analysis aims to understand public sentiments towards depression by analyzing tweets related to the topic. The following subsections provide a comprehensive overview of the research outcomes and the discussions surrounding them:

### 3.1. Implementation Requirements

#### 3.1.1 Data Requirements

In this project, the first step involves using the Snscrape library and setting search parameters to collect tweet data related to depression disorder. Subsequently, the collected tweet data is extracted using Snscrape and saved in JSON format, facilitating sentiment analysis in the next steps. To analyze the sentiment of the tweet data, a supervised learning approach is employed, enabling sentiment labeling for each tweet based on the available information. Thus, the project aims to gain a deeper understanding of sentiments surrounding the topic of depression disorder through the collected tweet data.

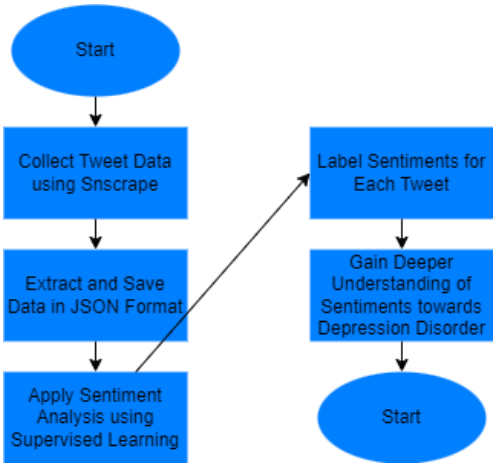


Figure 3. Data Requirements

### 3.1.2 Hardware and Software requirements

In this project, the hardware used is Asus UX305F with an Intel(R) Processor 5Y10 CPU @ 0.80GHz and 4.00 GB of RAM. The operating system utilized is Windows 10 Pro 64-bit, and several required software includes Google Colab, Visual Studio Code, Xampp, and a web browser. The programming languages employed are Python, HTML, and JavaScript. With this combination of hardware and software, it is anticipated that the project will run smoothly for the data analysis tasks.

### 3.2. Model Development

### 3.2.1. Data Collection

In the data collection process using Snscrape, the keyword "depresi" was used to search for related tweets in the Indonesian language. The data collection was performed one by one with a date range from April 24th to May 7th, 2023. The results were saved in the .json format and later converted to .csv format. The total number of successfully collected data is 4,460 entries.

### 3.2.2. Data Filtering

The data filtering process was conducted to select 1498 tweet data specifically discussing depression disorder from the total of 4,460 collected tweets. Filtering was crucial to focus on relevant data and obtain in-depth and accurate insights into Twitter users' sentiments about depression disorder. Removing irrelevant data ensures accurate analysis and a proper understanding of the public's perspectives on this issue.

### 3.2.2. Data Labelling

Table 1. Data labeling results

| Username | Tweet | Label |
|---|---|---|
| andrefeoh | Makin kesini kayanya makin banyak bgt orang depresi bahkan gila. | negative |
| itstikakid | @tionovita @audee84 @tanyakanrl Bangsal jiwa utk pasien depresi itu menyenangkan kata temen yg pernah ranap, isinya main & dikasih aktivitas yg fun. Plz ya gaes buat kelen yg kepikiran s word, lebih baik ke RSJ aja so they can take care of you | positive |
| napuspita | Gue punya depresi dan belajar untuk memahami hal2 yang mentrigger gue supaya gue bisa berfungsi baik ketika kerja & seminim mungkin nyusahin rekan kerja | netral |

### 3.2.2. Text Preprocessing

Table 2. Text preprocessing result

| Initial Tweet Data | Case Folding | Cleansing | Tokenizing | Stopword Removal | Stemming |
|---|---|---|---|---|---|
| Kata mereka aku moodbooster banget kalo diajak ngobrol, belum tahu ya kalo aslinya depresi berat. 😭🙏 | kata mereka aku moodbooster banget kalo diajak ngobrol, belum tahu ya kalo aslinya depresi berat. 😭🙏 | kata mereka aku moodbooster banget kalo diajak ngobrol belum tahu ya kalo aslinya depresi berat | ['kata', 'mereka', 'aku', ‘moodbooster', 'banget', 'kalo', 'diajak', 'ngobrol', 'belum', 'tahu', 'ya', 'kalo', 'aslinya', 'depresi', 'berat'] | moodbooster banget kalo diajak ngobrol ya kalo aslinya depresi berat | moodbooster banget kalo ajak ngobrol ya kalo asli depresi berat |

### 3.2.2. Exploratory Data Analysis (EDA)

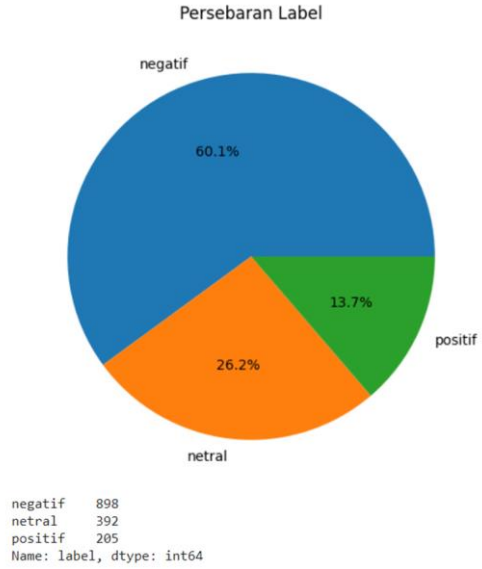


Figure 4. The results of the distribution of labels with a pie chart

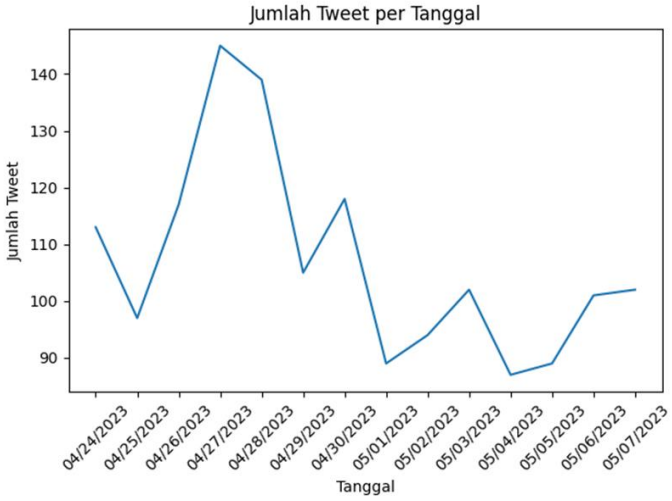


Figure 5. The results of the graph plotting tweets

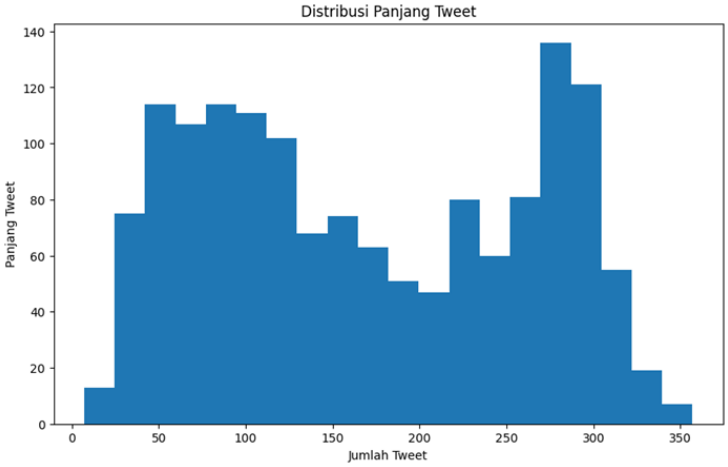


Figure 6. The graph results are the number of tweets in length

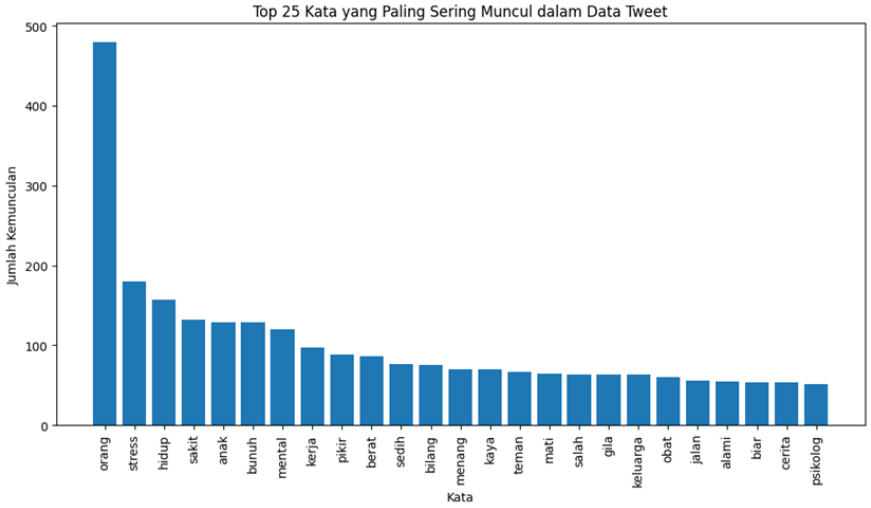


Figure 7. The graph results of word frequency used


Figure 8. Word cloud result

### 3.2.2. Data Splitting

In this research, the data was divided using the holdout method, with 80% of the data used as training data and 20% as testing data. The training data (1198) was used to train the model, while the testing data (300) was used to evaluate the model's performance and ensure accurate prediction capability on new, unseen data. This approach helps the model become more generalized and capable of providing accurate predictions on data it has not encountered before. By using a separate testing dataset, the research ensures a robust evaluation of the model's effectiveness and its ability to generalize well to real-world data scenarios.

### 3.2.2. TF-IDF Weighting

In the text weighting stage, the TF-IDF (Term Frequency-Inverse Document Frequency) method is used to calculate the weights of words in the text. The TF-IDF matrix reflects the numerical importance of words in the text. With this representation, we can identify words with high impact in the overall text analysis.

### 3.2.2. Naive Bayes Classification

Table 3. Classification results

| Classification Models | Accuracy | Processing Time |
|---|---|---|
| Multinomial NB | 0.9013 | 0.0963 |
| Bernoulli NB | 0.8536 | 0.0830 |
| Gaussian NB | 0.8837 | 0.1641 |

### 3.2.2. Classification Model Evaluation

| Classification Model Experiment | | Accuracy | Processing Time | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|---|---|---|
| Multinomial NB | Negative | 90,13% | 0.174 | 88% | 97% | 93% | 710 |
| | Netral | | | 93% | 76% | 84% | 328 |
| | Positive | | | 94% | 86% | 90% | 158 |
| Bernoulli NB | Negative | 85,36% | 0.078 | 82% | 98% | 89% | 710 |
| | Netral | | | 93% | 69% | 79% | 328 |
| | Positive | | | 91% | 64% | 75% | 158 |
| Gaussian NB | Negative | 88,37% | 0.076 | 100% | 84% | 91% | 710 |
| | Netral | | | 86% | 92% | 89% | 328 |
| | Positive | | | 64% | 100% | 78% | 158 |

### 3.3. Visualization

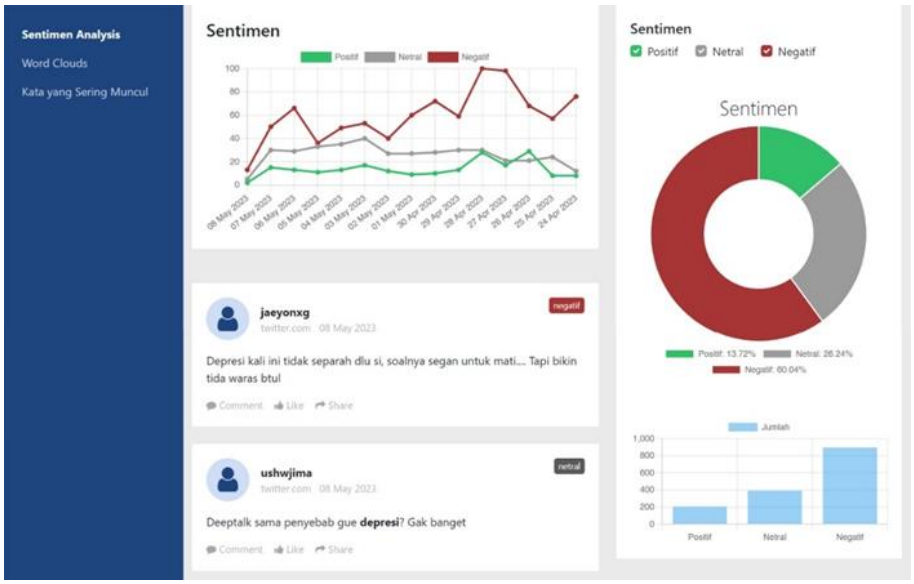


Figure 9. Visual Sentiment Analysis Page Display


Figure 10. Word Clouds Visualization Page Display


Figure 11. Depression Trends Visualization Page Display

## 4. CONCLUSION

The stages of developing a Twitter sentiment analysis classification model related to depression disorder encompass literature review, needs analysis, data collection, filtering, data labeling, and the TF-IDF weighting process. Three variations of the Naïve Bayes algorithm (Multinomial, Bernoulli, and Gaussian) were tested during model development. The results indicated that the Multinomial Naïve Bayes model achieved the highest accuracy (90.13%), followed by Gaussian Naïve Bayes (88.37%) and Bernoulli Naïve Bayes (85.36%).

Additionally, the Multinomial Naïve Bayes model exhibited favorable F1-Score evaluations for each sentiment class. Subsequently, this model was utilized for visualizing the sentiments of tweets related to depression disorder, presenting daily tweet count graphs, frequently occurring words, and word clouds for each sentiment category. This research provides a deeper understanding of sentiment analysis on Twitter concerning depression disorder and serves as a guide in building effective sentiment classification models.

## REFERENCES


[1] Basuma, D. (2013). Pencarian Alamat Fasilitas Umum Menggunakan Metode Vector Space Model (Studi Kasus Kota Pekanbaru) (Doctoral dissertation, UNIVERSITAS ISLAM NEGERI SULTAN SYARIEF KASIM RIAU).

[2] Budiman, K., Zaatsiyah, N., Niswah, U., & Faizi, F. M. N. (2020). Analysis of sexual harassment tweet sentiment on twitter in Indonesia using naïve Bayes method through national institute of standard and technology digital forensic acquisition approach. Journal of Advances in Information Systems and Technology, 2(2), 21-30.

[3] Chandra, D. N., Indrawan, G., & Sukaraja, I. N. (2016). Klasifikasi Berita Lokal Radar Malang Menggunakan Metode Naïve Bayes Dengan Fitur N-Gram. Jurnal Ilmiah Teknologi Informasi Asia, 10(1), 11-19.

[4] Fahrudin, T. M., Ruhui, A., Sari, F., Iffadah, A.Priambodo, J. (2018). Pendeteksian plagiarisme menggunakan algoritma Rabin-Karp dengan metode Rolling Hash. Jurnal Informatika Universitas Pamulang, 3(1), 39-45. S., Windyadari, C. C., & Khusnul, G. (2022). Pemodelan Teks Tweet pada Isu Pelecehan Seksual Berbasis Analisis Sentimen dan Leksikon Emosi. 2022(Senada), 12–23..

[5] Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge university press.Ratnawati, F. (2018). Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter. INOVTEK Polbeng-Seri Informatika, 3(1), 50-59

[6] A. D. Dwivedi, G. Srivastava, S. Dhar, and R. Singh, "A decentralized privacy-preserving healthcare blockchain for IoT," *Sensors (Switzerland)*, vol. 19, no. 2, pp. 1–17, 2019, doi: 10.3390/s19020326.

[7] Hakimi, F. D. D. (2018). Sistem analisis sentimen publik tentang opini pemilihan Kepala Daerah Jawa Timur 2018 pada dokumen twitter menggunakan naive bayes classifier (Doctoral dissertation, UIN Sunan Ampel Surabaya).

[8] Hidayatullah, M., Alam, S., & Jaelani, I. (2021). Sentiment Analysis of Police Performance On Twitter Users Using Naïve Bayes Method. RISTEC: Research in Information Systems and Technology, 2(2), 29-40.

[9] Hidayatullah, A. F., & Azhari, A. S. (2015, July). Analisis sentimen dan klasifikasi kategori terhadap tokoh publik pada twitter. In Seminar Nasional Informatika (SEMNASIF) (Vol. 1, No. 1).

[10] Indrayuni, E., & Wahyudi, M. (2015). Penerapan character N-gram untuk sentiment analysis review hotel menggunakan algoritma naive bayes. Konferensi Nasional Ilmu Pengetahuan dan Teknologi, 1(1), 83-88.

[11] Krisnanto, F., Tristiyanto, T., & Ardiansyah, A. (2018). Simulasi Sistem Informasi Komoditas Pasar Berbasis Web Menggunakan Metode Continuous Double Auction. Jurnal Komputasi, 6(2), 88-96.

[12] Lestari, A. R. T., Perdana, R. S., & Fauzi, M. A. (2017). Analisis sentimen tentang opini pilkada dki 2017 pada dokumen twitter berbahasa indonesia menggunakan naive bayes dan pembobotan emoji. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN, 2548, 964X

[13] Madani, S. A., Kazmi, J., & Mahlknecht, S. (2010). Wireless sensor networks: modeling and simulation. Discret. Event Simulations, (2004), 1-16.

[14] Mahardhika, Y. S., & Zuliarso, E. (2018). Analisis Sentimen Terhadap Pemerintahan Joko Widodo Pada Media Sosial Twitter Menggunakan Algoritma Naives Bayes Classifier.

[15] Manning, C. D. (2008). Introduction to information retrieval. Syngress Publishing,.

[16] Nurhuda, F., Sihwi, S. W., & Doewes, A. (2016). Analisis sentimen masyarakat terhadap calon Presiden Indonesia 2014 berdasarkan opini dari Twitter menggunakan metode Naive Bayes Classifier. ITSmart: Jurnal Teknologi dan Informasi, 2(2), 35-42..

[17] Novantirani, A., Sabariah, M. K., & Effendy, V. (2015). Analisis Sentimen pada Twitter untuk Mengenai Penggunaan Transportasi Umum Darat Dalam Kota dengan Metode Support Vector Machine. eProceedings of Engineering, 2(1).

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

[19] Priambodo, J. (2018). Pendeteksian plagiarisme menggunakan algoritma Rabin-Karp dengan metode Rolling Hash. Jurnal Informatika Universitas Pamulang, 3(1), 39-45.

[20] Ratnawati, F. (2018). Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter. INOVTEK Polbeng-Seri Informatika, 3(1), 50-59.

[21] Rini, D. C., Farida, Y., & Puspitasari, D. (2016). Klasifikasi menggunakan metode hybrid bayessian-neural network: studi kasus identifikasi virus komputer. Jurnal Matematika MANTIK, 1(2), 38-43.

[22] M. Aqib, R. Mehmood, A. Alzahrani, I. Katib, A. Albeshri, and S. M. Altowaijri, *Smarter traffic prediction using big data, in-memory computing, deep learning and gpus*, vol. 19, no. 9. 2019.

[23] Samsir, S., Ambiyar, A., Verawardina, U., Edi, F., & Watrianthos, R. (2021). Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes. Jurnal Media Informatika Budidarma, 5(1), 157-163.

[24] Subari, & Ferdinandus. (2015). Sistem Information Retrieval Layanan Kesehatan Untuk Berobat Dengan Metode Vector Space Model (Vsm) Berbasis Webgis. Snatika2, 3(November), 202–212.

[25] Tripathi, G., & Naganna, S. (2015). Feature selection and classification approach for sentiment analysis. Machine Learning and Applications: An International Journal, 2(2), 1-16.