E-ISSN : 2988-585X (Online)

Journal of Elektronik Sistem InformasI (JESII)

Volume 2 No 1 June 2024

DOI : 10.31848/jesii.v2i1.3416

Comparison Of The Results Of The Naïve Bayes Method And Synthetic Minority Over Sampling Technique In Sentiment Analysis Of User Reviews

Ilham Satria Al Munawar [1], Erwin Teguh Arujisaputra [2]

[1,2]Department of Information System, Univ. Kebangsaan Republik Indonesia, Indonesia

Article Info

Article history:

Received May 31, 24
Revised June 22, 24
Accepted Jun 24, 24

Keywords:

Naïve Bayes
Sentiment Analysis
SMOTE
User Review

ABSTRACT

The effectiveness of the Synthetic Minority Over-Sampling Technique (SMOTE) and the Naïve Bayes approach in sentiment analysis of user reviews is compared in this study. We examine if SMOTE, in contrast to the standard Naïve Bayes model, enhances sentiment classification accuracy by producing synthetic examples for the minority class, hence addressing class imbalance. The results will clarify how useful these methods are for sentiment analysis tasks, especially when working with unbalanced datasets. The rapid development of information technology encourages a variety of applications available on the Google Play Store. with various application categories, such as business, communication, education etc. One example of such an application is an online course application, namely Skill Academy from Ruang guru, which offers a variety of online guidance in the fields of education, self-development, and career. So, from this sentiment analysis will be carried out to understand a person's opinion and attitude towards a particular subject, theme, or entity in a text on the Skill Academy application from Ruang guru. This research aims to compare the performance of the Naive Bayes classification algorithm with the Synthetic Minority Over-Sampling Technique (SMOTE) on the sentiment of the Skill Academy application. This study shows the results of calculations without SMOTE and compares them with the results of SMOTE calculations. The results of this study are also expected to provide a better understanding of analyzing dataset imbalance on sentiment analysis results using Naïve Bayes and SMOTE techniques. In addition, it can be used in conjunction with appropriate evaluation methods to produce a more accurate model.

Corresponding Author:

Ilham Satria Al Munawar,
Information System Department, Faculty of Computer Science and Information Systems,
Univ. Kebangsaan Republik Indonesia.
Jln. Terusan Halimun No.37 (Pelajar Pejuang 45) Bandung, Jawa Barat, Indonesia. 40614
Email: ilham.sa@ukri.ac.id

1. INTRODUCTION

A wide range of applications with different application categories, such business, communication, education, etc., are available on the Google Play Store as a result of the information technology industry's increasingly rapid development [1]. An instance of such an application is the Ruang guru online education platform called Skill Academy. An online learning platform called Skill Academy provides a range of online coaching related to education, professional development, and personal growth. User reviews hold a wealth of sentiment data, but accurately analyzing this data can be challenging. This research explores the effectiveness of two techniques for sentiment analysis: the Naïve Bayes method, known for its efficiency, and Synthetic Minority Over-Sampling Technique (SMOTE), which tackles imbalanced datasets. By comparing their performance on user reviews, we aim to identify which approach delivers the most accurate sentiment classification [2].

An application's rating and review function on the Google Play Store allows users to rate applications on a scale of 1 to 5 and to share their experiences with them. Everyone has their own beliefs and viewpoints, and the launch of different programs at Skill Academy invites user feedback and opinions of all kinds. Sentiment analysis is the method of determining if a text contains favorable, negative, or neutral reviews and opinions. Sentiment analysis has several applications, such as monitoring social media, gauging public sentiment, and raising consumer satisfaction levels.

This study compared the performance of two methods in the analysis of user reviews sentiment: Naïve Bayes method and Synthetic Minority Over-Sampling Technique (SMOTE). Naive Bayes is known to be efficient, while SMOTE deals with data class imbalances. This study investigates whether SMOTE, which synthesizes samples for minority classes, can improve the accuracy of sentimental classification compared to the basic Naïve Bayes model [3]. Through this comparison, the study is expected to provide insight into the effectiveness of both techniques, in the task of sentimental analysis facing data imbalances.

The Naive Bayes method is one approach that can be utilized to sentiment analysis. Using the Bayes theorem, the Naive Bayes technique is a probabilistic classification system that forecasts the kind of data. Several benefits of the Naive Bayes approach include its ease of use, effectiveness, and capacity to handle unstructured data. Sentiment analysis suffers from an uneven dataset, meaning that the review data that is removed or taken does not equally represent positive and negative sentiment. As a result, sentiment classification performance suffers from a bias in the machine learning algorithm favoring the majority group.

Using the Synthetic Minority Over Sampling Technique (SMOTE) method is one way to refute this claim. To balance uneven data, the SMOTE approach employs oversampling. Through the creation of synthetic data from minority data, this approach operates.

2. METHOD

The research method used to analyze the sentiment of Skill Academy application user reviews with the naive Bayes method and Synthetic Minority Over-Sampling Technique (SMOTE) is divided into 2 stages, namely creating sentiment data and processing sentiment data.

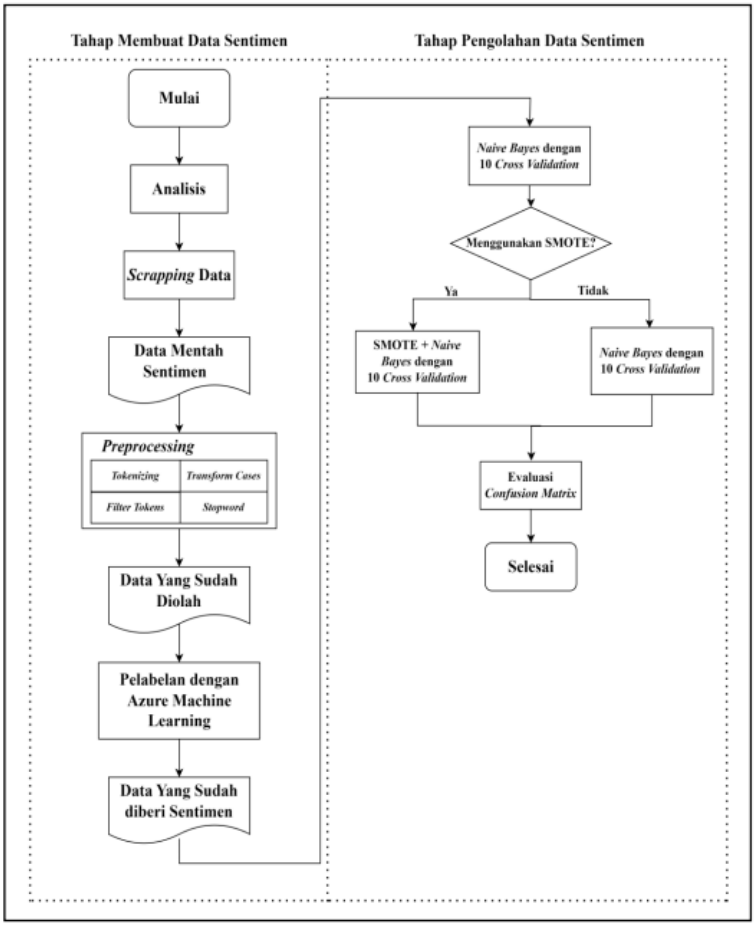

Figure 1. Research Methodology

2.1. Creating Sentiment Data

2.1.1. Analysis

The analysis aims to find out the needs needed in this research. Sentiment analysis is a subset of text mining that aims to understand, extract, and process text data to obtain information about the public's perspective or subjectivity towards an event, problem, or topic of conversation. Perspectives can include the author's perception of what they are writing about, their feelings while writing, or the emotional communication impact they want to have on the reader [1] .

Sentiment analysis is used to extract and evaluate sentiments or opinions contained in text such as product or service reviews. This technique can help producers or consumers in improving the quality of products or services. This technique can be used to identify positive, negative, or neutral sentiments in text [2].

2.1.2. Scrapping Data

Web Scraping is the process of extracting specific information from web pages using specific techniques to produce data that can be analyzed and used for various purposes. This technique is performed using computer code or specialized tools designed to retrieve specific data such as text, images, tables, or other types of information from the web page. Data scraping concentrates more on the extraction of the specific information required [3].

Web scraping technique is basically the process of copying data from a web but it is done automatically and in an organized manner. This technique collects data that is available on the internet, not from hacking that takes data directly from the web owner's server or database. Therefore, both the web owner

and the person doing the scraping benefit: the web owner gets traffic that can increase his valuation, and the person doing the scraping gets the necessary data.

2.1.3. Processing Data

Data preprocessing is the initial stage in data analysis that aims to clean and prepare data before further analysis is carried out. Data preprocessing improves the data to make it cleaner and more accurate before processing. Data preprocessing in this case plays an important role in improving the data to make it cleaner and more accurate before processing [4]. The following are the steps in data preprocessing based on information found in several studies:

1. Tokenizing: This stage aims to break the text into chunks of words and remove punctuation and numbers that make the next step easier.
2. Transform case: This stage is used to change uppercase letters to lowercase letters, and vice versa.
3. Stemming: This stage aims to convert the words in the data into their basic form. For example, the words "food" and "eat" will be converted into the base word "eat".
4. Token Filter (by Length): This stage removes words that are too short and too long, a minimum of 4 letters and a maximum of 25 letters.
5. Stop-word removal: This stage aims to remove words that have no meaning or are irrelevant in data analysis, such as common words like "and", "or", or "which".

After preprocessing is completed, the data is ready for further processing using appropriate data analysis techniques. Good and correct data preprocessing can help improve data quality and accuracy.

2.1.4. Labeling with Azure Machine Learning

Sentiment data labeling is the process of assigning sentiment labels to text data. The labeling includes positive, negative, or neutral. Sentiment data labeling is an important step in sentiment analysis, which is the process of extracting opinions, judgments, attitudes, and emotions from text data [5].

Azure Machine Learning is also a cloud service that makes it easy for developers to create, deploy and manage applications in the Microsoft environment. This service also helps Microsoft measure the value that users add to its cloud business on social media. This information can be used to develop new services that better suit user needs and improve customer satisfaction. Azure ML Studio includes data preprocessing, exploration, validation of modeling results, and method selection. It supports around other techniques such as: regression, anomaly detection, classification (binary and multi-class), and text analysis [6].

2.2. Sentiment Data Processing

2.2.1. Classify Naïve Bayes with 10 Cross Validation

Bayes theorem is the basic concept used by Naive Bayes, which involves calculating probability values to predict the class or label of certain data. In the Naive Bayes method, the class or label probabilities are based on the probabilities of the features present in the data. The prior probability is the initial probability of each class, while the probabilities are the feature probabilities of each class [7]. Cross Validation is used to estimate model performance on previously unseen data. Cross Validation divides the training data into several parts called folds. The model is trained on some folds and tested on other folds. This process is repeated several times so that each fold is used once as test
data [8].

2.2.2. Confusion Matrix Evaluation

Confusion matrix is a matrix used to evaluate the process of a classification model in terms of the number of correct and incorrect test data. Through this matrix we can determine the quality of the classification model performance [9]. The evaluation includes:

1. Accuracy refers to the ratio value of data detected by the test.

2. Precision refers to the value of the system's accuracy of the system information that displays the correct positive and negative data.
3. Recall refers to the value that indicates the degree of success to retrieve information regarding true positive and negative data. Recall is generated by comparing the true positive value with the number of true positive values.

3. RESULT AND DISCUSSION

In this section, it is explained the results of research and at the same time is given the comprehensive discussion. Results can be presented in figures, graphs, tables and others that make the reader understand easily [14], [15]. The discussion can be made in several sub-sections.

3.1. Creating Sentiment Data

3.1.1. Analysis

The analysis starts with problem identification, which is the basis of the problem understanding process. The author found that the development of online courses in Indonesia is increasing. User reviews are also an important part of business continuity to maintain the quality of a company's business or product. Therefore, this research focuses on the Skill Academy application from Ruang guru, by understanding the positive and negative evaluations of the application, to understand whether the services provided match the user's needs.
This final project analysis is carried out using 2 software, namely web harvy and RapidMiner. web harvy is used to perform web scrapping, which is useful for data mining. RapidMiner is used to analyze the data obtained.

3.1.2. Web Scrapping

After conducting the analysis stage, the process continues with web scraping (web mining). This stage is carried out to retrieve data from the sources used in this study. The tool used uses web harry software, by retrieving data on the play.google.com website with a sentiment time span between January and December 2022. Scraping is done by retrieving user sentiment data on the Skill Academy application, in the form of names, dates, ratings, and user comments.
After the above process is run, the data is retrieved and the results of the data obtained as from mining are downloaded in the form of excel files (.xls). The data collected amounted to 4,388 application user sentiments. After that, the data is processed through the preprocessing stage.

3.1.3. Data Preprocessing

The next process is data preprocessing, where the scraping results in an unstructured format are processed into basic words. This process is done through RapidMiner as shown in Figure 5.

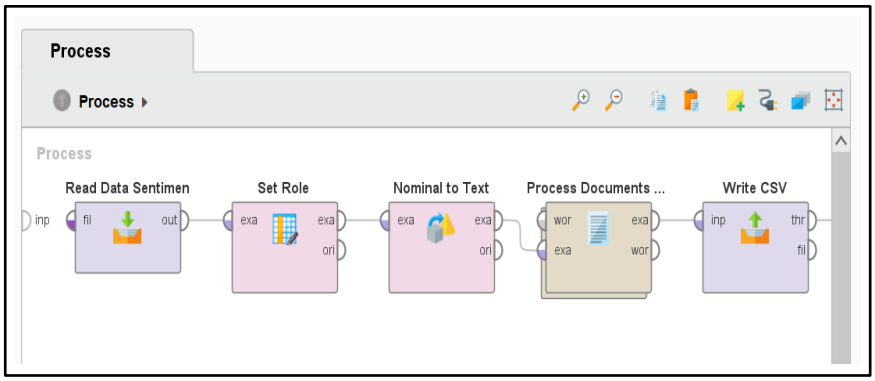

Figure 2. Sentiment Data Processing in Rapidminer Tool

The first step is to manually import the labeled data into RapidMiner for preprocessing. Rapidminer reads CSV or XLS files containing numerical or text data and converts them into a SampleSet format that can be used in data analysis. Set Role used to annotates the attributes, which will be used as labels and

and the next Type Nominal to Text, aims to convert nominal values into text.
Then the process document performs text data processing in which there are tokenize, transform cases, Filter Tokens and Filter Stopwords subprocesses.

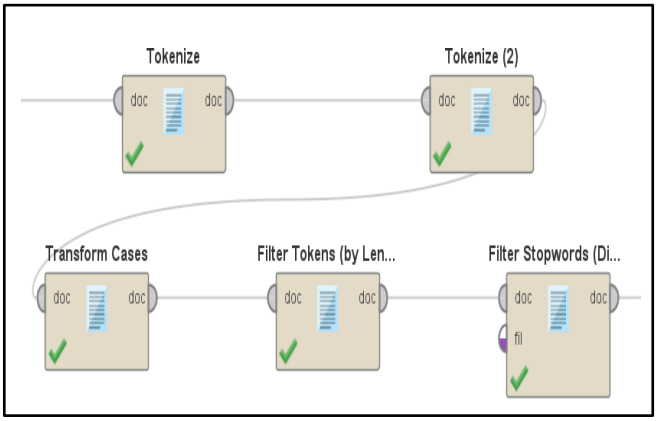

Figure 3. Sentiment Document Process

In Figure 3, the process in the sentiment document is explained, the steps are:

1. Tokenize
   The first tokenization process is the process of removing symbols that are not needed in the data. While the second tokenization process is the process of removing numbers to break sentences into word.
2. Transform Cases
   This process performs letter equalization from capital letters to lowercase letters.
3. Filter Tokens (by Length)
   This stage is used to limit the minimum and maximum characters in each word.
4. Filter Stop word (Dictionary)
   Removing words that have no meaning when standing alone, for example conjunctions. For this operator, there is no stop word in Indonesian, so the author must create a stop word text.

After that saving the preprocessing result data into CSV format

3.1.4. Labeling Sentiment

After doing the preprocessing process, the next process is the labeling process, which is the positive and negative labels on the sentiment. Before going through the labeling process, the mined data is then reviewed and moved to a different Microsoft Excel so that the labeling process can be done easily. After that, labeling is done using the Azure Machine Learning tool which utilizes the Add-ins feature in Microsoft Excel as an additional feature that can be installed in Microsoft Excel to expand its functionality [10]. Then attach the label name to the existing comment data, and the labels to be given are positive and negative which can be seen in Figure 4.

	Sentimen	Label
2	sangat bagus	positif
3	mantap keren	positif
4	bagus	positif
5	menambah ilmu	positif
6	sangat membantu	positif
7	mantap	positif
8	sangat bagus	positif
9	setalah pending sekrng kode voucher tidak terdaftar sangat	netral
10	parah marketing whatsapp skill academy tidak informatif sa	negatif
11	penipu bonus saya cair katanya cuma sampai sekarang mas	negatif
12	server video begitu buruk padahal jaringan internet stabil p	negatif
13	luar biasa materi pelajaran sangat mudah dipahami	positif
14	kelas kelasnya berkualitas materi tersampaikan	netral
15	saya sangat kecewa dengan event bonus career mentoring j	negatif
16	sistem cashbacknya menipu kita beli dengan harga ribu imir	negatif
17	materi pelatihan skill academy sangat bagus mudah fahami	positif
18	sangat bagus	positif
19	banyak merugikan	positif

Figure 4. Labeling Results Using Azure Machine Learning

The labeling also resulted in a total of 3558 positive sentiments, 277 neutral sentiments and 408 negative sentiments. In this study, both positive and negative sentiments are considered. Positive and negative sentiment classifications are generally preferred over neutral for sentiment analysis because they are easier to understand, interpret and apply. These classifications provide richer information about opinions and feelings, and are easier to learn and model with machine learning algorithms

3.2. Sentiment Data Classification Process

This sentiment data classification process uses naive Bayes with SMOTE and considers the process using naive Bayes without SMOTE. From different processes, it produces 2 different results to consider the results of the process using SMOTE and without using SMOTE in the classification results using the naive Bayes algorithm.

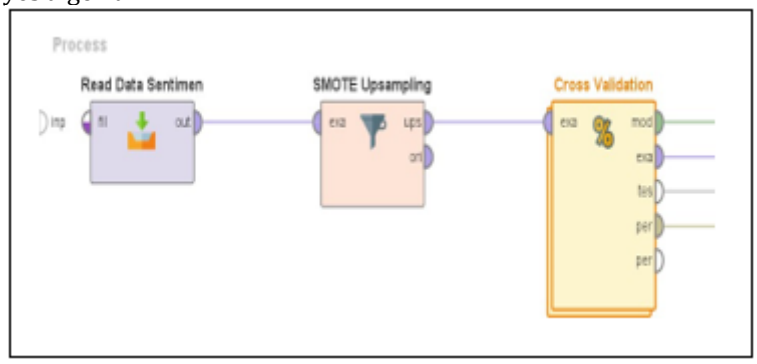

Figure 5. Sentiment Data Processing with SMOTE

Processing sentiment data that has been labeled starts with entering sentiment data. This research considers the use of SMOTE Up sampling and without SMOTE to determine the effectiveness of using SMOTE in balancing data classes. Then the data is classified using the naive Bayes algorithm on the Cross Validation operator.

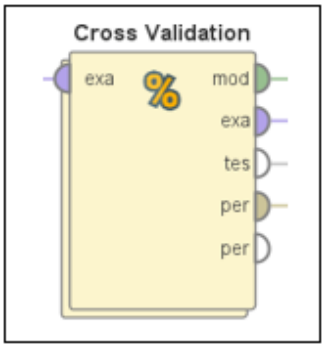

Figure 6. Cross Validation Operator

This stage performs operator validation using 10-fold cross validation which is useful for evaluating the performance of the naive Bayes model. In validation there are two processes, namely the training process and the testing process. In 10 cross validation operators, the data will be divided into 10 subsets. Each subset is then used as testing data once, while the rest is used as training data. This process is repeated 10 times, so that each data will be used as testing data once.

Inside the validation operator there are several processes that are used to generate the value of the evaluation. Processes in the validation operator validation operator in rapid miner can be seen in the following picture 7.

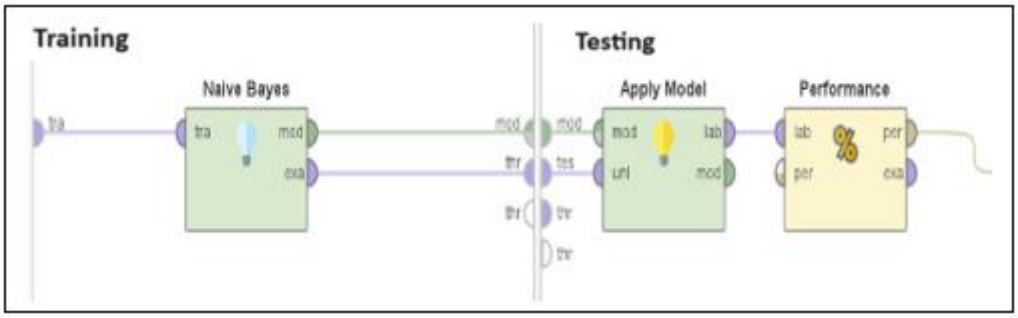

Figure 7. The Training Process and The Testing Process in Cross Validation

. Training process there is a classification algorithm that will be applied, namely naive Bayes, while in the testing process there is an Apply Model operator which is useful for running the naive Bayes model and a Performance operator which is useful for measuring the performance of the naive Bayes model itself.

3.3. Confusion Matrix Evaluation

Evaluation of confusion matrix results from the application of Naive Bayes algorithm in sentiment analysis using RapidMiner tool based on the evaluation of Accuracy, Precision, Recall and F-Measure values. The evaluation compares the results with SMOTE and with SMOTE. The results can be seen in the following table:

Table 1. The Result with SMOTE and without SOTE

Parameter	Naïve Bayes & SMOTE	Naïve Bayes
Accuracy	91,58%	82,86%
Precision	85,84%	79,92%
Recall	99,60%	99,25%
F-Measure	92,21%	87,06%

From the table above, it can be seen that the confusion matrix evaluation has increased when using SMOTE compared to without SMOTE.

4. CONCLUSION

Comparison of Confusion Matrix calculation results on sentiment data processing using Naive Bayes algorithm with SMOTE technique is greater and better than the classification results without using SMOTE technique. SMOTE technique can help improve model performance in cases of class imbalance in the data. this research compared the performance of the Naïve Bayes method and Synthetic Minority Over-Sampling Technique (SMOTE) for sentiment analysis of user reviews. We investigated whether SMOTE, which addresses class imbalance, could improve sentiment classification accuracy compared to the baseline Naïve Bayes model. The findings will reveal (1) if SMOTE is effective in enhancing sentiment analysis results for user reviews, and (2) the overall suitability of each technique for this task, particularly considering factors like dataset balance and computational efficiency. This technique can also be used in conjunction with appropriate evaluation methods to produce more accurate models. Synthetic Minority Over-sampling Technique (SMOTE) can help balance data in data processing using the naive Bayes Algorithm using the RapidMiner tool by adding synthetic samples to the minority class

so that the number of samples in the minority class becomes balanced with the number of samples in the majority class. The results obtained using the SMOTE technique can help improve model performance in cases of class imbalance in the data. This technique can also be used in conjunction with appropriate evaluation methods to produce more accurate models in cases of class imbalance in the data.

ACKNOWLEDGEMENTS

Thank you to all parties who have supported this research so that it can be completed properly. Especially, for the parents who have always supported and guided me in this research. I`m very grateful.

REFERENCES

[1] Y. Saputra, N. I. Putri, E. S. Nurpajriah, D. Jaelani, and A. Hamdani, “Perencanaan Strategis Sistem Informasi untuk Mendukung Keputusan Organisasi dengan Ward dan Peppard,” vol. 6, no. 2, pp. 137–145, 2023.

[2] Y. Saputra, E. S. Nurpajriah, S. Kustinah, and N. I. Pratiwi, “Perancangan Strategis Sistem Informasi Financial Planning Management dengan Robo-Advisor,” vol. 6, no. 2, pp. 127–136, 2023.

[3] A. R. Atmadja, A. Rahmawati, C. N. Alam, P. Dauni, and Y. Saputra, “Sentiment Analysis on Tourism Place using Naive Bayes,” *Proceeding 2023 17th Int. Conf. Telecommun. Syst. Serv. Appl. TSSA 2023*, pp. 1–6, 2023, doi: 10.1109/TSSA59948.2023.10366891.

[4] A. a. S. A. F. Y. C. Hasya, "Analisis Sentimen Berbasis Aspek pada Review Female Daily Menggunakan TF-IDF dan Naı̈ve Bayes," *Jurnal Media Informatika Budidarma,* vol. 5(2), pp. 422-430,, 2021.

[5] K. Parneet, "Sentiment analysis using web scraping for live news data with machine learning algorithms," *Materials today: proceedings,* vol. 6(5), pp. 3333-3341, 2022.

[6] F. a. M. Rezza, "Analisis Sentimen Terhadap Ulasan Aplikasi Pejabat Pengelola Informasi dan Dokumentasi Kementerian Dalam Negeri Republik Indonesia di Google Playstore Menggunakan Metode Support Vector Machine," *Jurnal Teknologi dan Komunikasi Pemerintahan,* vol. 4(1), pp. 1-13, 2022.

[7] S. I. M. A. D. S. a. P. W. B. Shevira, "Pengaruh Kombinasi dan Urutan Pre-Processing pada Tweets Bahasa Indonesia," *JITTER-Jurnal Ilmiah Teknologi dan Komputer,* p. 3(2), 2022.

[8] D. C. R. D. D. F. K. R. D. a. R. N. N. P. G. Naraswati, "Analisis Sentimen Publik dari Twitter Tentang Kebijakan Penanganan Covid-19 di Indonesia dengan Naive Bayes Classification," *Sistemasi: Jurnal Sistem Informasi,* Vols. 10(1),, pp. 222-238, 2021.

[9] A. M. Learning, "What is Azure Machine Learning?," [Online]. Available: https://learn.microsoft.com/en-us/azure/machine-learning/overview-what-is-azure-machine-learning?view=azureml-api-2. [Accessed 2 April 2024].

[10] H. R. a. W. J. Pranoto, "Implementasi Teorema Naı̈ve Bayes Pada Prediksi Prestasi Mahasiswa," *Jurnal Rekayasa Teknologi Informasi (JURTI), vol. 5(1), pp. 10-16, 2021. ,* vol. 5(1), pp. 10-16, 2021.

[11] A. F. N. M. D. S. K. a. A. R. P. C. B. Sonjaya, "The Performance Comparison of Classification Algorithm in Order to Detecting Heart Disease," *INTERNAL (Information System Journal),* vol. 5(2), pp. 166-175, 2022.

[12] S. N. a. M. Jajuli, "Integrasi Naive Bayes Dengan Teknik Sampling Smote Untuk Menangani Data Tidak Seimbang," *Nuansa Informatika,* vol. 14(1), p. 2020, 34-37

[13] M. J. G. a. A. H. Notz, "Evaluation of Sentiment Databases: A Comparison of Sentiment Databases through Social Listening Statements and Azure Machine Learning Studio," *Proceedings of the 2019 3rd International Conference on E-Business and Internet,* pp. 8-12, 2019.

[14] Z. M. Ahmad and M. S. M. Sirojul, "PERANCANGAN WEB E-COMMERCE UMKM RESTORAN BAKSO AREMA MENGGUNAKAN FRAMEWORK LARAVEL," *Jurnal Teknologi Terpadu,* vol. 5, no. 1, pp. 26-33, Juli 2019.