Naive Bayes and Support Vector Machine Algorithm for Sentiment Analysis Opensea Mobile Application Users in Indonesia

Laurenzius Julio Anreaja [1], Norma Nobuala Harefa[2], Julius Galih Prima Negara [3], Venantius Nathan Hermanu Pribyantara[4], Agung Budi Prasetyo[5].
[1]Information Systems Study Program, Faculty of Industrial Technology, Atma Jaya University Yogyakarta
[2]Information Systems Study Program, Faculty of Industrial Technology, Atma Jaya University Yogyakarta
[3]Departement of Informatics, Atma Jaya University Yogyakarta
[4]Information Systems Study Program, Faculty of Industrial Technology, Atma Jaya University Yogyakarta
[5]Information Systems Study Program, Faculty of Industrial Technology, Atma Jaya University Yogyakarta
Email: [1]laurenziusjulio11@gmail.com, [2]normanharefa@gmail.com, [3]julius.galih@uajy.ac.id, [4]venantiusnathan@gmail.com, [5]Agungpraset54@gmail.com

Abstract −Opensea is an NFT buying and selling application-based platform that is booming in the community. One way to find out the public's perception of the Opensea application is by sentiment analysis, as done in this study. Data that is used is user review data for the Opensea application in the Indonesian play store. The sentiment analysis technique used is the Naïve Bayes Classifier and the Support Vector Machine (SVM) method. Both are used to compare public responses from sentiment analysis of reviewed data labeled as positive, negative, and neutral. Based on this study, it was found that the Naive Bayes algorithm gives the results that class precision is 87.31%, class recall is 71.02%, and accuracy is 89.81%. While the SVM algorithm gives the results that class precision is 94.23%, class recall 71.96%, and Accuracy 90.78%. It is concluded that the SVM algorithm has a better performance than the Naive Bayes algorithm.

Keywords – Opensea, NFT, Sentiment Analysis, Google play store, SVM, Naive Bayes

I. INTRODUCTION

Opensea application users began to boom in Indonesia in early 2021 because a student from Semarang City, Sultan Gustaf Al Ghozali, got 1.5 billion from the sale of selfie photos in the form of NFT entitled Ghozali Everyday on the OpenSea platform [1]. The OpenSea Marketplace is another place where collectibles sold for Decentraland can be found. The Decentraland collection is linked to NFT. An NFT is a digital token that functions as a digital certificate of ownership for a digital asset such as an artist's digital collection or image. [2].

OpenSea has recorded $20.37 billion in sales and has over 1.2 million active traders in its network [3].
We opted to collect sales data from OpenSea, the largest active NFT marketplace. Statistics show that OpenSea has accumulated over $20 billion in trade volume, boasting more than 1.2 million traders. The data set was built using the provided OpenSea API, where we made our queries against the Events endpoint [4].

Play Store is a digital content provider service owned by Google that provides various online product stores such as applications, games, movies or music, and books of various categories. Google Play Store can be accessed through the website, android application, and Google TV. In the Google Play Store application, there are several features, one of which is the rating and review feature from users of available applications or services. A review or review is a text or sentence that contains an assessment or comment on a person's work. The importance of these reviews is often used as a benchmark for an application, whether it is recommended or not for new users [5].

Sentiment analysis or opinion mining is the process of understanding, extracting, and processing textual data automatically to obtain sentiment information contained in an opinion sentence. Sentiment analysis is carried out to see opinions or opinion tendencies towards a problem or object by someone, whether they tend to have negative or positive views or opinions [6].

In this study, sentiment analysis was carried out to see reviews from users of the OpenSea application. These reviews could be put into three categories, namely positive, neutral and negative.

Many studies have used machine learning algorithms with support vector machines (SVM) and Naïve Bayes (NB) being the most commonly used. Naïve Bayes (NB) is a technique based on Bayes' theorem. The Naive Bayes algorithm assumes that the presence of certain features in a class does not correlate with the presence of other features. This model is easy to build and very useful for very large data sets. Despite its simplicity, Naive Bayes is known to outperform even the most complex classification methods [7].

Support Vector Machine (SVM) is a classification and regression method commonly used for linear and non-linear problems. It has the advantage of applying linear splits to high-dimensional non-linear input data, and this is achieved by using the required kernel functions. The effectiveness of the Support Vector Machine is strongly influenced by the type of kernel function selected and

JISA (Jurnal Informatika dan Sains) (e-ISSN: 2614-8404) is published by Program Studi Teknik Informatika, Universitas Trilogi

applied based on the characteristics of the data. Many studies have reported that the Support Vector Machine is the most accurate method for text classification [8].

In previous research regarding the analysis of sentiment on E-Wallet Review (OVO). This study uses 500 positive reviews and 500 negative reviews as training data. The results of this study indicate that the use of the Naive Bayes algorithm produces an accuracy value of 93.10 percent. In comparison, the research results from the SVM algorithm are 91.30 percent. Based on these results, the accuracy value generated by the Naive Bayes algorithm and SVM was found that SVM is the best algorithm for classifying [9]. Also, previous research regarding the sentiment analysis of the Indonesian Police Mobile Brigade Corps based on Twitter posts using the SVM and NB methods resulted in an accuracy value of 86.96% with the SVM approach, 86.96% precision value, and 86.96% recall value [10].

The purpose of this study is to predict sentiment labels on reviews from users of the OpenSea application on the Google Play Store using the Naïve Bayes method and Support Vector Machine as a classification model.

II. RESEARCH METHODOLOGY

The object of this research is the Indonesian people's tweets against the metaverse on Twitter social media. In this study, there are several steps taken in analyzing the sentiments of the Indonesian people towards metaverse technology. The steps taken in this research can be seen in figure 1.

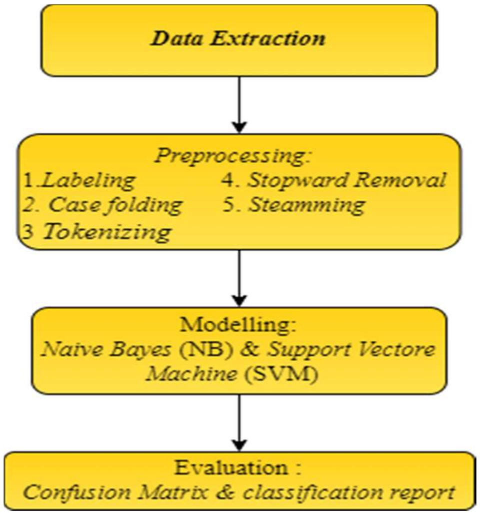


Figure 1. Research Process

2.1 Data Extraction

Collecting data in this study was obtained from reviews of users of the OpenSea application on the Google Play Store, using the web scraping technique, namely the technique which is used to extract data in large quantities large than the website where the data already extracted is saved in CSV (Comma Separated Value) format [11]. The web scrapper process in the google play store uses the Google-play-scrapper. Google-play-scrapper is Node js module to scrape application data from the Google Play store [12]. The data is then processed using the python language to go to the next stage, which is preprocessing.

2.2 Preprocessing Data

Pre-processing is a stage that is carried out to process and improve data so that it can be processed after the data to be analyzed has been obtained [13]. The following are:

1) Labeling the data in this study will be carried out by two people. The first person is tasked with manually classifying positive, negative, and neutral sentiments, while the second person re-examines the correctness of the classification results that have been carried out by the first person.

2) Case Folding is the stage to change sentences that have uppercase (capital letters) into lowercase (lowercase). This is done in order to obtain structured and consistent data in the use of capital letters.

3) Tokenizing is the stage to separate sentences into several pieces of words called tokens. Separate words using space punctuation restrictions. The following is an example of tokenizing in the table.

4) Stop word removal is a step to get rid of various useless words in a sentence with the help of the Sastrawi library. Sastrawi library is a library that can also be used to perform stopword removal with unimportant words in Indonesian [14].

5)Stemming is a step taken by researchers to remove prefixes and suffixes for each token with the help of Sastrawi stemmer. Sastrawi stemmer is a stemmer library that is used to overcome the problem of changing words with words into basic words in the Indonesian language [15].

2.3 Modeling

Modeling is a method in which a model represents correlation relationships between one set of data and the other set of data [16]. The first step in starting data modeling is to partition or divide the data into training data and testing. The data modeling process for the case of sentiment analysis is carried out using several classification methods, including supervised learning, such as Support Vector Machine (SVM) and Naïve Bayes. This algorithm was chosen because it is a commonly used method in sentiment analysis.

Naive Bayes Classifier is the probability used to determine text document class and can process large amounts of data with high accuracy results [17]. SVM (Support Vector Machine) is a machine learning algorithm that is used to divide each data class to find the most optimal hyperplane [18]. The SVM algorithm tries to find a hyperplane to maximize the distance between classes. In this way, SVM can guarantee the ability of high generalization for data that will be predicted [19].

2.4 Evaluation

The evaluation process in this research uses a confusion matrix and classification report. The confusion matrix is a table that is used to describe the performance of a classification algorithm. A confusion matrix visualizes and summarizes the classification algorithm's performance in label comparison and machine learning prediction results [20]. Classification reports are used to measure the predictive quality of the classification of a particular algorithm so as to show the precision, recall, and accuracy of an application of the model algorithm [21]. The aim is to see and compare the accuracy, precision, and recall of SVM and Naive Bayes models in analyzing sentiment.

III. Results And Discussion (10 pt, Capital, Bold)

3.1 Data Extraction

The dataset that will be used in the research is taken from user reviews of the Opensea application on the play store by doing web scrapping via google-play-scrapper and python language. The dataset collected reviews in Indonesian as many as 1028 reviews.

3.2 Preprocessing

3.2.1 Labelling

The dataset obtained is then carried out manually, labeling the sentiment by two people. Where the first person gives the label and the second person checks the correctness of the labeling. The results of the labeling obtained 731 positive sentiments, 231 negative sentiments, and 86 neutral sentiments. The results of the labeling stage can be seen in table 1.

Table 1. Labeling

Review	Sentimen
Mantap gw Dapet 3 ETH Berkat aplikasi ini	Positive
KOK NFT SAYA HILANG ATAU BERKURANG.. TIDAK ADA PENJELASAN DARI PIHAK OPENSEA	Negative

3.2.2 Case Folding

For the dataset that has gone through the labeling process then, every uppercase letter in the comments column will be changed to lowercase, and the number will be removed. The results of the case folding process can be seen in Table 2.

Table 2. Case Folding

Review	Sentimen
mantap gw dapet eth berkat aplikasi ini	Positive
kok nft saya hilang atau berkurang.. tidak ada penjelasan dari pihak opensea	Negative

3.2.3 Tokenizing

At this stage, the sentence is broken down into words with punctuation and whitespace boundaries. The results of the tokenizing process can be seen in Table 3.

Table 3. Tokenizing

Review	Sentimen
mantap gw dapet eth berkat aplikasi ini	Positive
kok nft saya hilang atau berkurang tidak ada penjelasan dari pihak opensea	Negative

3.2.4 Stopword Removal

After the dataset goes through the tokenizing process, the next step is to delete words that are not important and interfere with the sentiment analysis process through the Stopword Removal stage. The results of this stage can be seen in table 4.

Table 4. Stopword Removal

Review	Sentimen
mantap gw dapet eth berkat aplikasi	Positive
kok nft hilang berkurang penjelasan pihak opensea	Negative

3.2.5 Stemming

The data preprocessing process is then ended by removing the affixes for each word so that the resulting words are only the basic words. The results of this stage can be seen in Table 5.

Table 5. Stemming

Review	Sentimen
mantap gw dapet eth berkat aplikasi	Positive
kok nft hilang kurang penjelasan dari pihak opensea	Negative

3.3 Data Modelling

The training and testing data used in this data modeling is 80%: 20%. This means that from 1028 the training data collection owned is 822 records while the testing data owned is 206 records. Based on the results of the tests conducted on the Opensea application, user comment test data, which consists of 3 labels, namely positive, negative, and neutral, using the Naive Bayes classifier obtained a match accuracy with the train data of 89.81%. Meanwhile, using the Support Vector Machine algorithm, the accuracy was 90.78%. This means that the Naive Bayes model is more accurate than SVM in this study.

The visualization of the bar chart of the number of positive, negative, and neutral sentiments from the Support Vector Machine can be seen in table 8, and the distribution of the most dominant words in the positive, neutral, and negative labels are presented in the form of a word cloud. The word cloud in the positive class is shown in figure 9, while the negative class word cloud is shown in figure 10.

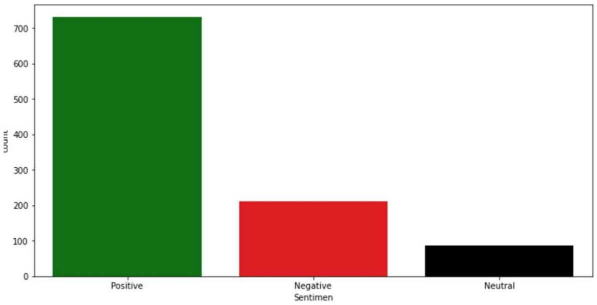


Figure 2. Distribution of the results of the analysis using the Support Vector Machine

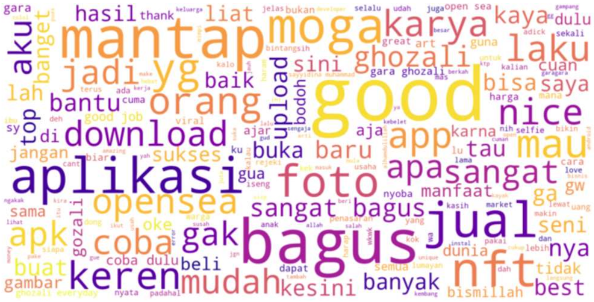


Figure 3. Positive words

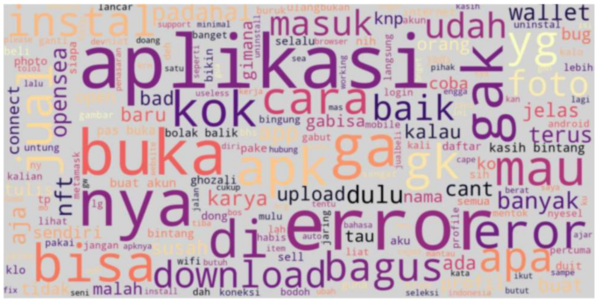


Figure 4. Negative Words

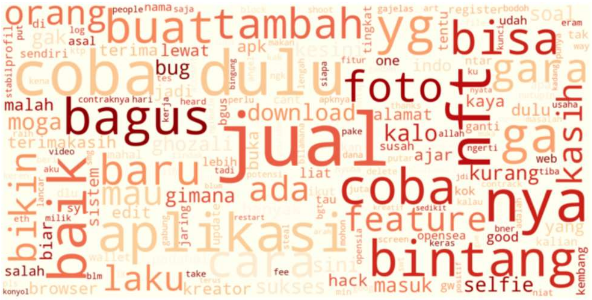


Figure 5. Neutral Words

3.4 Evaluation

After the model is created, it needs to be evaluated using a confusion matrix. Evaluation is done using confusion matrix so we can know the exact result of true positive, true negative, true neutral false positive, false negative, and false neutral. True positive, true negative, true neutral false positive, false negative, and false neutral. True positive is the successful positive class classified as the positive class, the true negative is the successful negative class classified as the negative class, and true neutral is the successful neutral class classified as the positive class. false positive is a negative class, and neutral class is classified as a positive class, false negative is a positive class, and neutral class is classified as a negative class. False neutral is a negative class, and a positive class is classified as a neutral class. The classification report is used to determine the class recall and class precision on a model that is being run.

In the evaluation of the naive bayes model with the confusion matrix, the results obtained the results of the true Positive = 143, false positive = 15, true negative = 37, false negative = 5, true neutral = 5, and false neutral = 1. The results of the confusion matrix can be seen in Figure 6.

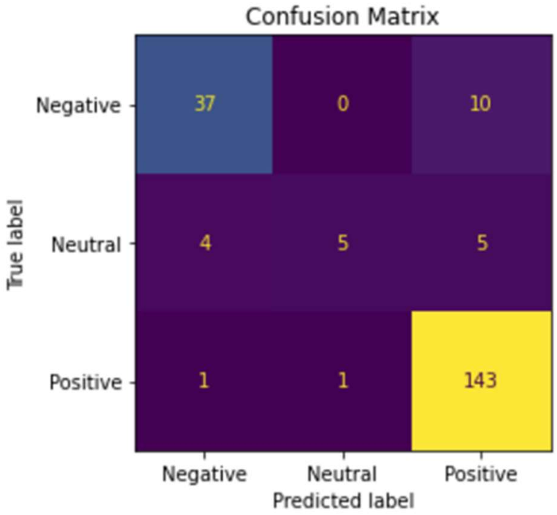
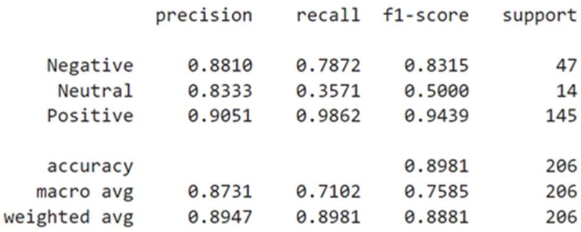


Figure 6. Confusion Matrix Naive Bayes

While the classification report in Naive Bayes shows class precision as negative, neutral, and positive is 88.10%, 83.33%, and 90.51%, while class recall is negative, positive, neutral that is 78.72%, 35.71%, 98.62%, so the results obtained are an average class precision of 87.31%, an average class recall of 71.02%, an accuracy of 89.81%. The results of the Classification report can be seen in figure 7.



```
              precision    recall  f1-score   support

    Negative     0.8810    0.7872    0.8315        47
     Neutral     0.8333    0.3571    0.5000        14
    Positive     0.9051    0.9862    0.9439       145

    accuracy                         0.8981       206
   macro avg     0.8731    0.7102    0.7585       206
weighted avg     0.8947    0.8981    0.8881       206
```

Figure 7. Classification Report Naive Bayes

Table 6 below is a combination of the results of the confusion matrix with the classification report on the Naive Bayes evaluation and is shown in tabular form so that it is easy to see the correlation.

Table 6. Summary of Confusion Matrix and Classification Report Naive Bayes

	True Negative	**True Neutral**	**True Positive**	**Precision**
Pred Negative	37	4	1	88.10%
Pred Neutral	0	5	1	83.33%
Pred Positive	10	5	143	90.51%
Recall	78.72%	35.71%	98.62%	

In the evaluation of the SVM model with the confusion matrix, the results of the true positive = 144, false positive = 16, true negative = 38, false negative = 3, true neutral = 5, and false neutral = 0. The results of the confusion matrix can be seen in figure 8.

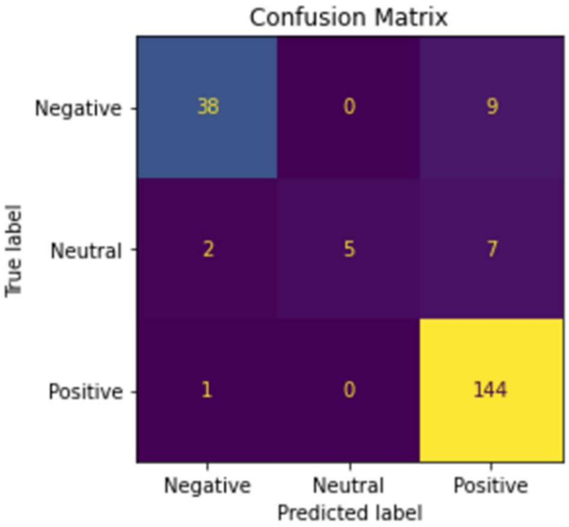
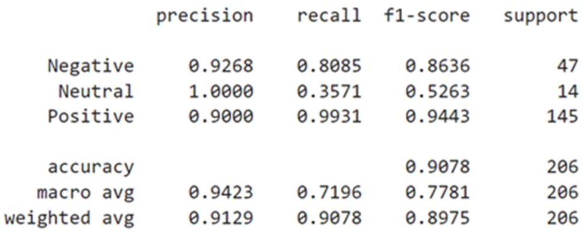


Figure 8. Confusion Matrix SVM

While the classification report in SVM shows Class precision as negative, neutral, and positive that is 92.68%, 100.00%, and 90.00%% while Class recall is negative, neutral, positive is 80.85%, 35.71%, 99.31%, so the results obtained are an average class precision of 94.23%, an average class recall of 71.96%, and an accuracy of 90.78%. The results of the Classification report can be seen in figure 9.



```
              precision    recall  f1-score   support

    Negative     0.9268    0.8085    0.8636        47
     Neutral     1.0000    0.3571    0.5263        14
    Positive     0.9000    0.9931    0.9443       145

    accuracy                         0.9078       206
   macro avg     0.9423    0.7196    0.7781       206
weighted avg     0.9129    0.9078    0.8975       206
```

Figure 9. Classification Report SVM

The combination of the results of the confusion matrix with the classification report on the Support Vector Machine evaluation is shown in tabular form so that it is easy to see the correlation. The result can be seen in table 7.

Table 7. Summary of Confusion Matrix and Classification Report SVM

	True Negative	**True Neutral**	**True Positive**	**Precision**

Pred Negative	38	2	1	92.68%
Pred Neutral	0	5	0	100.00%
Pred Positive	`7	9	144	90.00%
Recall	80.85%	35.71%	99.31%	

Based on the results of the comparison of the SVM and Naive Bayes algorithms, In table 8, the Naive Bayes algorithm gives the results that class precision is 87.311%, class recall is 71.02%, and accuracy is 89.81%. While the SVM algorithm gives the results that class precision is 94.23%, class recall 71.96%%, and accuracy 90.78%.The result can be seen in table 8.

Table 8. Performance comparison between Naive Bayes and SVM

	Accuracy	**Precision**	**Recall**
Naïve Bayes	89.81%	87.31%	71.02%
SVM	90.78%	94.23%	71.96%

IV. CONCLUSION

Based on the results of the sentiment analysis in this study, it can be seen that the Opensea Application Review dataset predicted using the Naïve Bayes algorithm and SVM showed significant results. The Naive Bayes algorithm gives the results that class precision is 87.31%, class recall is 71.02%, and accuracy is 89.81%. While the SVM algorithm gives the results that class precision is 94.23%, class recall 71.96%%, and accuracy 90.78%. It is concluded that the SVM algorithm has a better performance than the Naive Bayes algorithm. This research has not compared the performance with other machine learning algorithms besides Naive Bayes and SVM, so it is necessary to make a comparison with other classification machine learning algorithm models. Such as lexicon, linear regression, and random forest so that later it can improve the accuracy of sentiment classification in similar research.

REFERENCES

[1] Natasya Salim, "Minat Terhadap NFT Bertambah Sejak Nama Ghozali Viral, Pakar Serukan Adanya Regulasi," 2022. https://www.abc.net.au/indonesian/2022-01-19/minat-nft-di-indonesia-meningkat-tapi-waspadai-risiko-kejahatan/100765112.

[2] Mohamed-amine et all, "How should metaverse augment humans with disabilities ?," in *13th Augmented Human International Conference Proceedings*, 2022, p. 9, [Online]. Available: https://archive-ouverte.unige.ch/unige:160466.

[3] DeepRadar, "NFT MarketPlace Ranking," 2022. https://dappradar.com/nft/marketplaces (accessed Jul. 11, 2022).

[4] B. White, *Characterizing the OpenSea NFT Marketplace*, vol. 1, no. 1. Association for Computing Machinery, 2021.

[5] S. A. Aaputra, "Sentiment Analysis Analisis Sentimen E-Wallet Pada Google Play Menggunakan Algoritma Naive Bayes Berbasis Particle Swarm Optimization," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 3, no. 3, pp. 377–382, 2019.

[6] I. Rozi, S. Pramono, and E. Dahlan, "Implementasi Opinion Mining (Analisis Sentimen) Untuk Ekstraksi Data Opini Publik Pada Perguruan Tinggi," *J. EECCIS*, vol. 6, no. 1, pp. 37–43, 2012.

[7] D. D. Tran, T. T. S. Nguyen, and T. H. C. Dao, "Sentiment Analysis of Movie Reviews Using Machine Learning Techniques," *Lect. Notes Networks Syst.*, vol. 235, no. December 2017, pp. 361–369, 2022, doi: 10.1007/978-981-16-2377-6_34.

[8] I. Santoso, Windu Gata, and Atik Budi Paryanti, "Penggunaan Feature Selection di Algoritma Support Vector Machine untuk Sentimen Analisis Komisi Pemilihan Umum," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 3, no. 3, pp. 364–370, 2019, doi: 10.29207/resti.v3i3.1084.

[9] I. Ajzen, "The theory of planned behavior," *Organ. Behav. Hum. Decis. Process.*, vol. 50, no. 2, pp.

179–211, Dec. 1991, doi: 10.1016/0749-5978(91)90020-T.

[10] Sularso et al, "Sentiment Analysis of the Indonesian Police Mobile Brigade Corps Based on Twitter Posts Using the SVM And NB Methods," *J. Phys. Conf. Ser.*, vol. 1201, 2019, [Online]. Available: https://www.researchgate.net/publication/333585160_Sentiment_Analysis_of_the_Indonesian_Police_Mobile_Brigade_Corps_Based_on_Twitter_Posts_Using_the_SVM_And_NB_Methods/.

[11] R. Hanifah and I. S. Nurhasanah, "Implementasi Web Crawling Untuk Mengumpulkan Web Crawling Implementation for Collecting," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 5, pp. 531–536, 2018, doi: 10.25126/jtiik20185842.

[12] Latif et al, "Data Scraping from Google Play Store and Visualization of its Content for Analytics," 2019, [Online]. Available: https://www.researchgate.net/publication/342636207_Data_Scraping_from_Google_Play_Store_and_Visualization_of_its_Content_for_Analytics.

[13] D. P. Sari, "Pemanfaatan NFT Sebagai Peluang Bisnis Pada Era Metaverse," *J. Akrab Juara*, vol. 7, no. 1, pp. 237–245, 2022, [Online]. Available: https://dspace.uii.ac.id/handle/123456789/29069.

[14] B. Siswanto, "Sentiment Analysis in Indonesian on Jakarta Culinary as A Recommender System," 2021, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9702772.

[15] M. A. Rosid, A. S. Fitrani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, "Improving Text Preprocessing for Student Complaint Document Classification Using Sastrawi," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 874, no. 1, 2020, doi: 10.1088/1757-899X/874/1/012017.

[16] Will Koehrsen, "Modeling: Teaching a Machine Learning Algorithm to Deliver Business Value," 2018. https://towardsdatascience.com/modeling-teaching-a-machine-learning-algorithm-to-deliver-business-value-ad0205ca4c86 (accessed May 11, 2022).

[17] Pristiyono, M. Ritonga, M. A. Al Ihsan, A. Anjar, and F. H. Rambe, "Sentiment analysis of COVID-19 vaccine in Indonesia using Naïve Bayes Algorithm," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1088, no. 1, p. 012045, 2021, doi: 10.1088/1757-899x/1088/1/012045.

[18] M. Bloodgood, "Support Vector Machine Active Learning Algorithms with Query-by-Committee Versus Closest-to-Hyperplane Selection," *Proc. - 12th IEEE Int. Conf. Semant. Comput. ICSC 2018*, vol. 2018-Janua, no. 2, pp. 148–155, 2018, doi: 10.1109/ICSC.2018.00029.

[19] S. Fransiska and A. Irham Gufroni, "Sentiment Analysis Provider by.U on Google Play Store Reviews with TF-IDF and Support Vector Machine (SVM) Method," *Sci. J. Informatics*, vol. 7, no. 2, pp. 2407–7658, 2020, [Online]. Available: http://journal.unnes.ac.id/nju/index.php/sji.

[20] A. Kulkarni, D. Chong, and F. A. Batarseh, "Foundations of data imbalance and solutions for a data democracy," *Data Democr. Nexus Artif. Intell. Softw. Dev. Knowl. Eng.*, pp. 83–106, Jan. 2020, doi: 10.1016/B978-0-12-818366-3.00005-8.

[21] Muthukrisman, "Understanding the Classification report through sklearn," 2018. https://muthu.co/understanding-the-classification-report-in-sklearn/.