Jurnal Sisfotek Global
ISSN (Online): 2721 - 3161, ISSN (Print): 2088 – 1762
DOI: http://dx.doi.org/10.38101/sisfotek.v13i1.2943
Vol. 13, No. 1, March 2023, pp. 8-14

Analysis of Sentiment Adiraku App Reviews on Google Play Store Using Vector Machine Support Algorithm and Naïve Bayes

Bayu Padilah[1], Adi Rizky Pratama[2] & Ayu Ratna Juwita[3]

[1,2,3] Universitas Buana Perjuangan, Karawang, Indonesia, 41316
E-mail: [1]if19.bayupadilah@mhs.ubpkarawan.ac.id, [2]Adi.Rizky@ubpkarawang.ac.id, [3]ayurj@ubpkarawang.ac.id

ARTICLE HISTORY

Received : March 1st, 2023
Revised : March 26th, 2023
Accepted : March 30th, 2023

KEYWORD

Adiraku
Sentiment Analysis
Support Vector Machine
Naive Bayes

ABSTRACT

The Adiraku application is considered to be able to facilitate and facilitate customers so that there is no need to come to the branch office to get information related to the number of installments that must be paid, due dates, credit simulations, and Adira Finance information offers to customers. A large number of reviews from users received makes it difficult for developers to read them, it will take too much time and effort if they have to read and analyze them manually. To find out which reviews are classified as positive or negative reviews. need a sentiment analysis of the review. This study aims to find out how the opinions or opinions of its users on the services of the application, by analyzing these sentiments through a classification process using two algorithms, namely Support Vector Machine and Naïve Bayes. The data used amounted to 2000 data obtained from Google Playstore. Data is labeled into 2 classes namely positive class and negative. Furthermore, the data is divided into 70% training data and 30% testing data and methods used for testing using Bernoulli Naïve Bayes and Linear Kernel. It was concluded that the number of user reviews of the Adiraku application on the Google Play Store showed more positive comments, amounting to 1412 positive and negative reviews, which was 588 reviews. The Support Vector Machine algorithm performs better by getting the best accuracy value of 96%, while the Naïve Bayes algorithm gets an accuracy value of 85%.

1. Introduction

Currently, the development of information technology is very rapid and fast, including in Indonesia itself. With technology, it is basically to make it easier for humans to run things [1]. Adiraku is an information system that is still new. This information system was officially released on February 20, 2020, to facilitate and facilitate customers so that they do not need to come to the branch office to get information related to the number of installments that must be paid, due dates, credit simulations, and Adira Finance information offers to selected customers as a form of promotion. The Adiraku application is available for free download on both the Google Play store and App Store. The Google Play store displays a rating of 4.4 for the Adiraku application, with a total of 52,000 reviews in the comments section. These reviews include both positive and negative reviews such as complaints, criticisms, or suggestions [2].

The number of users who use the Adiraku App making reviews of the App is also increasing. However, because a large number of reviews from users received makes it difficult for developers to read them, it will take too much time and effort if they have to read and analyze them manually and this kind of method is not recommended because it is not effective. Meanwhile, these reviews can influence the Adiraku Application in making improvements to the application [3]. To find out whether a review is positive or negative, sentiment analysis of the review is required.

In a study of sentiment analysis on the Gojek application, [4] employed the Support Vector Machine and K Nearest Neighbor algorithms. The KNN method with k=22 achieved an accuracy of 82.14%, precision of 82.28%, and recall of 95.43%. On the other hand, the SVM method with linear kernel and C=1 parameters obtained accuracy, precision, and recall values of 87.98%, 88.55%, and 95.43%, respectively. Further research by [5] on Sentiment Analysis of Online News Media App Reviews On Google Play Using the Support Vector Machine and Naïve Bayes Algorithm Methods The results show that SVM (Support Vector Machines) is 88% superior to Naïve Bayes by 87% and is derived from the tendency of public opinion in Google Play about online news media applications skewed positively, from the number of positive opinions of 5160 while negative by 455. A

recent study on Sentiment Analysis in BCA Mobile App Reviews was conducted by [6], utilizing the BM25 method and the Improved K-Nearest Neighbor algorithm for document classification. The study evaluated the performance of the model using 5-fold testing, and obtained optimal results with a k-value of 10, achieving a precision value of 0.946, recall of 0.934, f-measure of 0.939, and accuracy of 0.942.

Further research by [7], SVM and Naive Bayes was used to analyze public sentiment regarding the KPK's Hand Capture Operations. Researchers used the SVM algorithm to measure the accuracy of social media in the analysis of KPK arrests. Each dataset consists of 78 positive tweets and 78 negative tweets related to KPK arrest operations. The results of research by Herawati showed an accuracy of 83.79% and AUC of 0.910 in using the SVM algorithm for the analysis. Other research conducted by [8] about Sentiment analysis of the impact of coronavirus on twitter using Naïve Bayes and SVM methods. The data used as many as 1104 data taken from tweets on Twitter using 3 sentiments, namely positive, negative, and neutral. In testing with the Naive Bayes Classifier algorithm produced a data accuracy value of 81.07% while the SVM algorithm produced a data accuracy value of 79.96%. Then, on the research [9] The results of research using the Support Vector Machine show that the accuracy of the classification of public sentiment on Twitter towards online loans is 62.00%. Despite this, the accuracy results are considered quite good. In addition, sentiment analysis using SVM successfully classified people's sentiment on Twitter towards online loans. From the results of the classification, it was found that negative sentiment dominated with a percentage of 59%, while positive sentiment only reached 41%.

Furthermore, research by [10] From the results of analysis and testing of Youtube comments regarding the Samsung Galaxy Z Flip 3 gadget with a total of 9,597 comments, it was found that users gave more positive opinions on design aspects and negative opinions on aspects of price, specifications, and brand image. This study used the CRISP-DM model and compared the classification methods of Naïve Bayes (NB), Support Vector Machine (SVM), and k-Nearest Neighbor (k-NN), and the results showed that the SVM classification model gave the best results. The average accuracy of SVM was 96.43% for the four aspects analyzed, namely the design aspect of 94.40%, the price aspect of 97.44%, the specification aspect of 96.22%, and the brand image aspect of 97.63%. Lastly, research by [11] about Twitter Sentiment Analysis Post-Covid-19 Online Lecture Using Support Vector Machine and Naïve Bayes Algorithms. This study used naïve bayes algorithm and support vector machine (SVM) to analyze sentiment with performance results obtained. The naïve bayes algorithm has an accuracy of 81.20%, a time of 9.00 seconds, a recall of 79.60%, and a precision of 79.40%. As for the SVM algorithm, it has an accuracy of 85%, a time of 31.60 seconds, a recall of 84%, and a precision of 83.60%. This performance was achieved in the first iteration for naïve bayes and the 423rd iteration for the SVM algorithm.

Based on the background above, this study was titled 'Sentiment Analysis of Adiraku App Reviews on the Google Play Store Using Vector Machine And Naïve Bayes Support Algorithms'. This study aims to find out how the opinions or opinions of its users on the services of the application, by analyzing these sentiments with a classification process using two, namely the Support Vector Machine and Naïve Bayes algorithms. In addition, this research was also carried out for the quality and service contained in the Adiraku application.

2. Research Method

The research method will be carried out there are several stages, starting from data collection, pre-processing, and classification using 2 algorithms, namely the Support Vector Machine and Naïve Bayes for the last stage, namely evaluation. More details can be seen in figure 1.

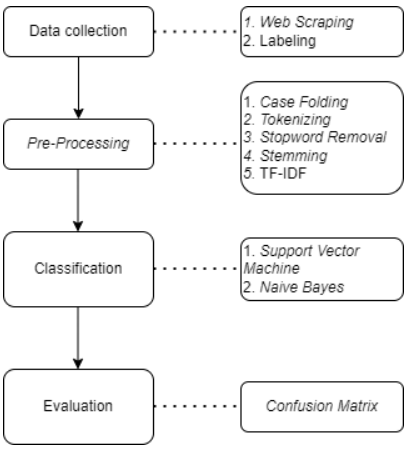

Figure 1. Flowchart Diagram

2.1. Data Collection

In this research, data were collected from user reviews of the Adiraku application on the Google Play store. The data was extracted from approximately 2000 recent user reviews to ensure the presence of active users, thus providing a more up-to-date representation of user opinions. This technique aimed to ensure that the comments were still relevant and indicative of the current user sentiment towards the Adiraku application. For the method of collecting data on the adiraku application on the Google Play store, namely Web Scraping. In this study, the tool used was Google Colab. Then the already obtained review data is stored in the form of .csv. From the dataset that has been obtained, there is some information, namely a username, score, date and time, and content.

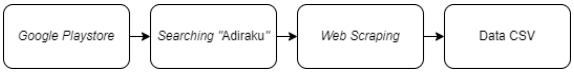

Figure 2. Data Collection

2.2. Pre-Processing

For the pre-processing stage, the raw data will be carried out through case folding, tokenizing, stemming, stopwords removal, and TF-IDF processes. Pre-processing stage to avoid the occurrence of imperfect random data or inconsistent data[12].

2.3. Classification

The study employed the Support Vector Machine and Naïve Bayes algorithms for classification. Before classification, the dataset was partitioned into two subsets, namely the training and testing data. The Support Vector Machine and Naïve Bayes models were constructed using the training data and subsequently assessed using the testing data. The performance and accuracy of the models were measured to determine their classification capability.

2.3.1. Support Vector Machine

Support Vector Machine is a classification algorithm that is carried out by determining the hyperplane. A good hyperplane will be right in the middle of the two classes, so it has the longest distance to the outermost data in both classes[13]. In this study, the authors used a linear kernel formula. Here's the linear kernel equation. It can be seen in Table 1.

Kernel	Equation
Linear	$K(x_i x) = x_i^T x$

Table 1. Kernel Formulas

2.3.2. *Naïve* Bayes

The classification of the Naïve Bayes method is a method that utilizes simple statistics based on the Bayes theorem that assumes the presence or absence of a class with other features. The Naïve Bayes method is used for classification processes to determine f1_score accuracy, recall, and precision [14]. In this study, researchers used Naive Bayes with the Bernoulli method following the calculation formula of Naïve Bayes Bernoulli.

$$P(C|X) = \frac{P(X|C)\,P(H)}{P(X)}$$

Information:

P(C|X) = Probability of class c if given x

P(X|C) = Probability of occurrence of x in class c

P(C) = Probability of occurrence of c in general

P(X) = Probability of occurrence of x in general

2.4. Evaluation

This stage is carried out making sure that the testing is correct. This evaluation is about finding the best results from the test results. To measure the level of accuracy against then using the confusion matrix, the calculation is with accuracy, precision, and recall[15].

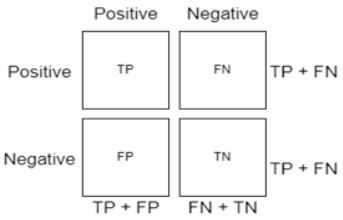

Figure 3. Confusion Matrix

Information:

TP (True Positive) = Positive data classified correctly

TN (True Negative) = Negative data classified correctly

FP (False Positive) = Negative data classified positively

FN (False Negative) = Positive data classified negatively

3. Result and Discussion

3.1. Data Collection

The data retrieval process in the existing Adiraku application reviews on the Google Play store uses the Web Scraping method using Google Colab for data obtained as much as 2000 data. The dataset can be seen in Figure 4.

	userName	score	at	content
0	sutrisno nasarudin	5	2023-01-31 05:40:52	Good job
1	Joker Pramz	5	2023-01-31 04:56:27	cepat pelayanannya
2	Joshua Sihombing	5	2023-01-31 02:18:52	Sangat membantu dalam melakukan pembayaran ang...
3	Dodi Sumirat	5	2023-01-30 22:13:19	Memberikan kemudahan , semoga aplikasi ya berg...
4	Ari Ehmpoey	1	2023-01-30 13:43:49	Lemot
...	...	...	...	...
1995	Ki Arya Sumeudang	5	2022-11-04 14:34:26	0k
1996	Ansar Gusasi	5	2022-11-04 11:20:08	Kenapa susah loginnya
1997	CI2X	1	2022-11-04 10:40:24	Aplikasi nggk guna klu sering eror" an Setiap ...
1998	Cristian Dwisanto	5	2022-11-04 09:47:01	Mantap deh adira
1999	Fleet Admiral	5	2022-11-04 06:21:23	Simple and fast

2000 rows × 4 columns

Figure 4. Dataset

After getting the dataset, the data is labeled into two classes, namely the Positive and Negative classes, where the Positive value has a score value of 3-5 while for Negative 1-2.

	content	score	Label
0	Mantap penanganan admin nya	1	negatif
1	banyak yg LBH mahal dri tmpt blanja online yg ...	3	positif
2	mantap	5	positif
3	Gila, belum transaksi saldo gopay tiba2 hilang...	1	negatif
4	Hari hari down mulu bang	1	negatif
5	Sangat membantu sekali	5	positif

Figure 5. Labeling Dataset

3.2. Pre-Processing

After going through the data collection process, the next process is to clean the data by pre-processing. Here are the pre-processing stages of this study.

a. Case Folding

This stage changes the data from uppercase to lowercase, removes numbers, punctuation, and blank characters in the document, or removes emoticons on the data. It can be seen in Table 2.

Before	After
Banyak promo menarik, thanks ADIRAKU	banyak promo menarik, thanks adiraku

Table 2. Case Folding Results

b. Tokenizing

The next stage is Tokenizing. At this stage, the breakdown of each word that was previously in the form of a sentence, document, or paragraph into certain parts based on each word. It can be seen in Table 3.

Before	After
cepat pelayanannya	[‘cepat’, ‘pelayanannya’]

Table 3. Tokenizing Results

c. Stopword Removal

The next step is Stopword Removal. This stage performs word filtering or separates unnecessary words in the data. It can be seen in Table 4.

Before	After
saya sudah bayar angsuran 11, tapi sisa	[‘saya’], [‘sudah’], [‘bayar’], [‘tapi’], [‘sisa’]

Table 4. Stopword Removal Results

d. Stemming

The next stage is stemming, where this process is to remove all the affixes that can be made to the word and turn it into a base word. Stemming is also done to reduce the variation of the same basic word. It can be seen in Table 5.

Before	After
Bagus keren mantul dah ngga	[‘bagus’, ‘keren’, ‘mantul’]

Table 5. Stemming Results

e. TF-IDF

The process of determining the significance of a successfully extracted word is known as weighting. This stage aims to assign a weight to each word, which will be utilized as a feature. The quantity of data or documents to be processed directly influences the number of features, with a larger dataset leading to more features. It can be seen in Figure 6.

```
(0, 2108)     0.2591044675539185
(0, 2024)     0.16602556076744804
(0, 1781)     0.2506745994633166
(0, 1703)     0.14499951933645663
(0, 1590)     0.23796342518038754
(0, 1510)     0.2284693238846259
(0, 1384)     0.13061477959901066
(0, 1264)     0.23291890554650366
(0, 1218)     0.2852900295613335
(0, 1004)     0.15940673915814282
(0, 767)      0.18054778153167383
(0, 542)      0.2852900295613335
(0, 285)      0.22088843040850972
(0, 219)      0.518208935107837
(0, 205)      0.17405566025505262
(0, 7)        0.26997245772668715
(1, 873)      0.591087633051676
(1, 3)        0.8066073456480342
(2, 629)      1.0
```

Figure 6. TF-IDF Results

3.3. Classification

At the data classification stage taking as many as 2000 data then the data is labeled into 2 classes, namely Positive and Negative. This classification process uses Google Colab with Python programming language and uses Support Vector Machine and Naïve Bayes algorithms. The results of the comparison can be seen in Figure 7.

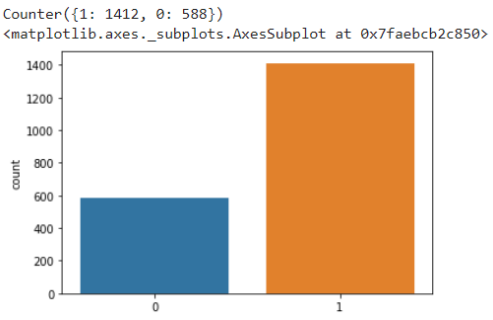

Figure 7. Negative and Positive Sentiment Classification Results

The picture above shows that the value 1 is interpreted as the Positive class and the value 0 it is interpreted as the Negative class for the Positive class it gets a total of 1412 data while the Negative class is 588 data. Furthermore, the data is divided into training data and testing data with a ratio of 0.3. Which means that the data in the training data amounts to 70%, which is 1,400 data, and for testing data which amounts to 30 %, which amounts to 600 data. The data that will be used in the classification to find the best accuracy value using the Support Vector Machine and Naïve Bayes algorithms is testing data containing user reviews of the Adiraku application.

3.3.1. Support Vector Machine Algorithm

The first algorithm used to classify user review data of the Adiraku Application is the Support Vector Machine with a linear kernel function to determine the accuracy level of the processed data. can be seen in the following table.

Algorithm	Accuracy	Precision	Recall
Support Vector Machine	96%	0.96%	0.99%

Table 6. Classification Results for SVM Kernels

Based on table 6 above, the *Support Vector Machine* Algorithm gets 96% accuracy results, 0.96% Precision, and 0.98% *Recall*.

3.3.2. Naïve Bayes Algorithm

The second algorithm used to classify user review data of the Adiraku Application is Naïve Bayes with the Bernoulli method to find out the accuracy level of the processed data. can be seen in the following table.

Algorithm	Accuracy	Precision	Recall
Naïve Bayes	85%	0.85%	0.97%

Table 7. Classification Results for Naïve Bayes

Naïve Bayes algorithm gains 85% accuracy, 0.8 5% Precision, and 0.97% Recall. The results that have been obtained from the two algorithms, will then be compared in the next discussion.

3.3.3. Comparison of Support Vector Machine and Naïve Bayes

After obtaining the results of the level of accuracy in the two algorithms, the next step is to compare them. Beris a comparison between the Support Vector Machine and Naïve Bayes algorithms Can be seen in figure 8 below.

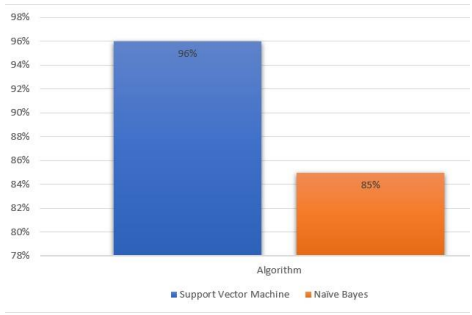

Figure 8. SVM and Naïve Bayes Algorithm Accuracy Value Comparison Chart

The chart presented above allows us to conclude that the Support Vector Machine algorithm exhibits superior performance in classifying sentiment analysis on user reviews of the Adiraku application on the Google Play store compared to the Naïve Bayes method.

3.4. Evaluation

This final stage aims to recalculate the results of the accuracy values obtained by each algorithm using the Confusion Matrix.

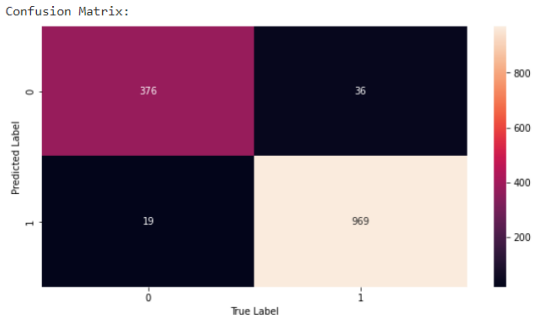

Figure 9. Confusion Matrix Support Vector Machine

Based on figure 9 above, it can be seen that the number of TP is 376, FP is 19, FN is 36, and TN is 969. To calculate the manual is with TP+FP+FN+TN = 1400, Next (TP+TN)/1400 =0.960. The result of the Confusion Matrix of the Support Vector Machine algorithm is 96%.

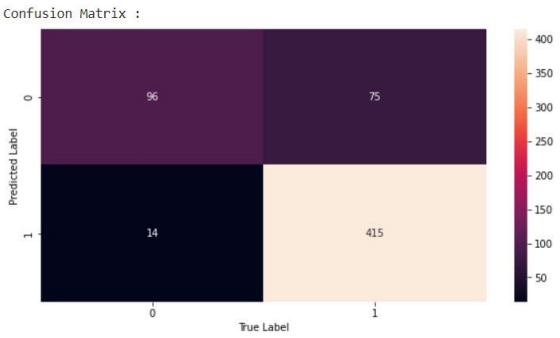

Figure 10. Confusion Matrix Naïve Bayes

Figure 10 above can be seen that the number of TP is 96, FP is 14, FN is 75, and TN is 415. To calculate the manual is with TP+FP+FN+TN = 600, Next

(TP+TN)/600 = 0.851. The result of the Confusion Matrix of the Naïve Bayes algorithm is 85%.

4. Conclusion

Based on the test results, it can be concluded that the number of user reviews of the Adiraku application on the Google Play Store shows a greater number of positive comments, namely 1412 positive reviews compared to the number of negative reviews, which is 588 negative reviews out of a total of 2000 reviews taken by the Web Scraping method. From the test results, it was found that the Support Vector Machine algorithm has a better performance value than the Naïve Bayes algorithm. Evidenced by the accuracy value obtained by the Support Vector Machine algorithm got an accuracy value of 96%, while the Naïve Bayes algorithm got an accuracy value of 85%. The amount of data used in the classification is 2000 data from Web Scraping results on the Google Play store, next the data is divided into 2, namely training data of 1400 data and testing data of 600 data. It can be concluded that the Support Vector Machine algorithm has a much better accuracy value for classification in this study compared to Naïve Bayes.

References

[1] L. Y. Siregar and M. I. P. Nasution, "Perkembangan Teknologi Informasi Terhadap Peningkatan Bisnis Online," *HIRARKI J. Ilm. Manaj. dan Bisnis*, vol. 02, no. 01, pp. 71–75, 2020, [Online]. Available: http://journal.upp.ac.id/index.php/Hirarki%0APERKEMBANGAN

[2] "FINANCE CABANG KOTA SALATIGA DENGAN PENDEKATAN TUGAS AKHIR Diajukan Kepada Program Studi Akuntansi Untuk Memperoleh Gelar Sarjana Akuntansi Oleh : SAMUEL PUTRA YOFINDA," 2020.

[3] M. D. Hendriyanto, A. A. Ridha, and U. Enri, "Analisis Sentimen Ulasan Aplikasi Mola Pada Google Play Store Menggunakan Algoritma Support Vector Machine," *INTECOMS J. Inf. Technol. Comput. Sci.*, vol. 5, no. 1, pp. 1–7, 2022, doi: 10.31539/intecoms.v5i1.3708.

[4] M. N. Muttaqin and I. Kharisudin, "Analisis Sentimen Pada Ulasan Aplikasi Gojek Menggunakan Metode Support Vector Machine dan K Nearest Neighbor," *UNNES J. Math.*, vol. 10, no. 2, pp. 22–27, 2021, [Online]. Available: http://journal.unnes.ac.id/sju/index.php/ujm

[5] U. Kusnia and F. Kurniawan, "Analisis Sentimen Review Aplikasi Media Berita OnlinePada Google Playmenggunakan Metode Algoritma Support Vector Machines (SVM) Dan Naive Bayes," *Explor. IT*, vol. 14, no. 36, pp. 24–28, 2022.

[6] P. P. A. Indriya Dewi Onantya, Indriati, "Analisis Sentimen Pada Ulasan Aplikasi BCA Mobile Menggunakan BM25 Dan Improved K-Nearest Neighbor," *J-Ptiik.Ub.Ac.Id*, vol. 3, no. 3, pp. 2575–2580, 2019, [Online]. Available: http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/4754

[7] H. Hernawati and W. G. Kedua, "Sentimen Analisis Operasi Tangkap Tangan Kpk Menurut Masyarakat Menggunakan Algoritma Support Vechtor Machine, Naïve Bayes, Berbasis Particle Swarm Optimizition," *Fakt. Exacta*, vol. 12, no. 3, p. 230, 2019, doi: 10.30998/faktorexacta.v12i3.4992.

[8] C. F. Hasri and D. Alita, "Penerapan Metode Naïve Bayes Classifier Dan Support Vector Machine Pada Analisis Sentimen Terhadap Dampak Virus Corona Di Twitter," *J. Inform. dan Rekayasa Perangkat Lunak*, vol. 3, no. 2, pp. 145–160, 2022, [Online]. Available: http://jim.teknokrat.ac.id/index.php/informatika

[9] D. S. Utami and A. Erfina, "Analisis Sentimen Pinjaman Online di Twitter Menggunakan Algoritma Support Vector Machine (SVM)," *SISMATIK (Seminar Nas. Sist. Inf. dan Manaj. Inform.*, vol. 1, no. 1, pp. 299–305, 2021.

[10] J. W. Iskandar and Y. Nataliani, "Perbandingan Naïve Bayes, SVM, dan k-NN untuk Analisis Sentimen Gadget Berbasis Aspek," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 6, pp. 1120–1126, 2021, doi: 10.29207/resti.v5i6.3588.

[11] H. Setiawan, E. Utami, and S. Sudarmawan, "Analisis Sentimen Twitter Kuliah Online Pasca Covid-19 Menggunakan Algoritma Support Vector Machine dan Naive Bayes," *J. Komtika (Komputasi dan Inform.*, vol. 5, no. 1, pp. 43–51, 2021, doi: 10.31603/komtika.v5i1.5189.

[12] I. P. Rahayu, A. Fauzi, and J. Indra, "Analisis Sentimen Terhadap Program Kampus Merdeka Menggunakan Naive Bayes Dan Support Vector Machine," vol. 4, pp. 296–301, 2022, doi: 10.30865/json.v4i2.5381.

[13] M. I. Putri and I. Kharisudin, "Penerapan Synthetic Minority Oversampling Technique (SMOTE) Terhadap Analisis Sentimen Data Review Pengguna Aplikasi Marketplace Tokopedia," *Prism. Pros. Semin. Nas. Mat.*, vol. 5, pp. 759–766, 2022, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/prisma/

[14] M. T. Anjasmoros, I. Istiadi, and F. Marisa, "Analisis Sentimen Aplikasi Go-Jek Menggunakan Metode SVM Dan NBC (Studi

Kasus: Komentar Pada Play Store)," *Conf. Innov. Appl. Sci. Technol. (CIASTECH 2020)*, no. Ciastech, pp. 489–498, 2020.

[15] A. S. Rahayu and A. Fauzi, "Komparasi Algoritma Naïve Bayes Dan Support Vector Machine (SVM) Pada Analisis Sentimen Spotify," vol. 4, pp. 349–354, 2022, doi: 10.30865/json.v4i2.5398.