JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. IX No 4, September 2023, hlm. 665-672 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v9i4.2539
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

COMPARISON OF NBC, SVM, KNN CLASSIFICATION RESULTS IN SENTIMENT ANALYSIS OF MOBILE JKN

Nadya Bethry Balqies Tjikdaphia[1*], Sulastri[1]
[1]Fakultas Teknologi Informasi dan Industri, Universitas Stikubank
email: [*]nadyabethrybalqiestjikdaphia@mhs.unisbank.ac.id

Abstract: The JKN Mobile application is a mobile application created to facilitate healthcare administration in Indonesia since 2017. The application has been downloaded by over 10 million users and has received 484,000 diverse reviews, including positive, negative, and neutral feedback. The average rating given by users is 4.5 out of 5 stars. This research aims to perform sentiment analysis on user reviews found in the Google Play Store review column. The methods used for sentiment analysis are Naive Bayes, K-Nearest Neighbor (K-NN), and Support Vector Machine (SVM). The test results show that with a 10% test data and 90% training data proportion, the SVM method achieves the highest accuracy of 95%. Naive Bayes follows with an accuracy of 87%, and K-NN with an accuracy of 75%.

Keywords: JKN mobile application, sentiment analysis, naive bayes, k-nearest neighbor (K-NN), support vector machine (SVM).

Abstrak: Aplikasi Mobile JKN adalah sebuah aplikasi yang dibuat untuk mempermudah administrasi kesehatan di Indonesia sejak tahun 2017. Aplikasi ini telah diunduh lebih dari 10 juta pengguna dengan 484 ribu ulasan beragam positif, negatif, dan netral. Rata-rata rating yang diberikan pengguna adalah 4,5 bintang dari 5 bintang. Penelitian ini bertujuan untuk melakukan analisis sentimen terhadap ulasan pengguna yang terdapat di kolom *review Google Play Store*. Metode yang digunakan untuk analisis sentimen adalah Naive Bayes, K-Nearest Neighbor (K-NN), dan Support Vector Machine (SVM). Hasil pengujian menunjukkan bahwa dengan menggunakan proporsi data uji sebesar 10% dan data *training* sebesar 90%, metode SVM mencapai akurasi tertinggi sebesar 95%. Diikuti oleh Naive Bayes dengan akurasi 87%, dan K-NN dengan akurasi 75%.

Kata kunci: JKN mobile, analisis sentimen, naïve bayes, k-nearest neighbor (K-NN), support vector machine (SVM).

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. IX No 4, September 2023, hlm. 665-672 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v9i4.2539
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

INTRODUCTION

In this era of globalization, BPJS Kesehatan has developed a mobile application called JKN Mobile to facilitate users in health registration and management. This application has been introduced since 2017 [1] and has more than 10 million users with an average rating of 4.5 stars out of 5 stars and a total of 513 thousand reviews [2]. The aim of this research is to analyze the sentiment of JKN Mobile application reviews on the Google Play Store using the Naive Bayes, K-Nearest Neighbor (K-NN), and Support Vector Machine (SVM) methods.

Naive Bayes classification is one of the classification algorithms based on Bayes' theorem with the assumption of independence between the existing features. This algorithm is very popular because it is simple, efficient in resource usage, and able to provide good results in many applications [3]. Support Vector Machine (SVM) is a machine learning algorithm used for classification and regression tasks. SVM works by building a model that can separate two classes by finding the best hyperplane that has the maximum margin between the classes [4]. K-Nearest Neighbor (K-NN) is a machine learning algorithm used for classification and regression tasks. KNN works by calculating the distance between the data to be classified and the existing training data. KNN is very simple and easy to understand but may face challenges in handling high-dimensional data [5].

The data used is the user review dataset of JKN Mobile taken from the Google Play Store. The analysis stage involves data preprocessing, such as text cleaning and extraction of relevant features. In the data retrieval process, the number of data ratings 1-5 is unbalanced, so the SMOTE step needs to be performed. SMOTE works by creating new synthetic data from the minority class through interpolation between existing data points. The synthetic data generated by SMOTE is in the feature space, not in the data space, because the scraping results are unbalanced [6]. Furthermore, the Naive Bayes, K-NN, and SVM methods are applied to the dataset. Evaluation is done using evaluation metrics such as accuracy, precision, recall, and F1-score.

The test results show that SVM has the highest accuracy of 85.71%, followed by Naive Bayes with 76.70% accuracy, and K-NN with 52.74% accuracy. Previous studies have also tested these three algorithms with similar results [7]. In another study, 5390 data were used for training and 599 data for testing. The results obtained from this research indicate that the accuracy of the K-Nearest Neighbor method is 54%, while the accuracy of the Support Vector Machine method with kernel is 79% [8]. There is also a similar study with accuracy results of 92.50% for Naive Bayes, 93.00% for SVM, and 95.00% for K-NN [9]. In this study, a single SVM algorithm was used with an accuracy of 82.20% [10].

METHOD

Research Methodology

Knowledge Discovery on Database (KDD) is considered a suitable method for this research, referring to the process of extracting useful knowledge from data in a database. This method consists of four main stages: data collection, data preprocessing, data mining, and evaluation [11].

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. IX No 4, September 2023, hlm. 665-672 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v9i4.2539
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

Data Collection

Data collection was conducted by filtering user reviews of the JKN Mobile application from the Google Play Store using Google Colab. These reviews have been labeled as positive, negative, or neutral based on their content.

The number of negative reviews with a rating of 1-2 is 858, the number of neutral reviews with a rating of 3 is 80, and the number of positive reviews with a rating of 4-5 is 262. The total collected data is 1200. Since the data appears to be imbalanced, a SMOTE step is needed to balance the data used.

```
[ ] df_busu['score'].value_counts()

1    749
5    211
2    109
3     80
4     51
Name: score, dtype: int64
```

Image 1. Data Collection Results

Preprocessing Data

Data preprocessing is an important step in preparing data before performing further analysis or processing. Its purpose is to clean, organize, and transform raw data into a more structured form that is easily understood and suitable for the analysis or modeling needs.

Several commonly used methods in data preprocessing are case folding, tokenizing, filtering, stopwords removal, stemming, and term weighting using TF-IDF [9].

Case Folding

Case Folding is the process of converting all text characters to either lowercase or uppercase to avoid irrelevant differences in writing.

	sentiment	content
0	NEGATIF	tengah malam di buat login error mulu padahal ...
1	NETRAL	baru mau daftar online semua data sudah saya i...
2	NEGATIF	jujur ini apk nya sedikit nyusahin karena lagi...
3	NEGATIF	daftar baru pakai nomer telfon 8org beda kok g...
4	NEGATIF	kenapa setelah update gak bisa di buka klo se...

Image 2. Case Floding Results

Tokenizing

Tokenizing is the process of separating text into smaller units, such as words or tokens, to facilitate analysis.

	sentiment	content
0	NEGATIF	[tengah, malam, di, buat, login, error, mulu, ...
1	NETRAL	[baru, mau, daftar, online, semua, data, sudah...
2	NEGATIF	[jujur, ini, apk, nya, sedikit, nyusahin, kare...
3	NEGATIF	[daftar, baru, pakai, nomer, telfon, 8org, bed...
4	NEGATIF	[kenapa, setelah, update, gak, bisa, di, buka,...

Image 3. Tokenizing *Results*

Filtering

Filtering involves removing irrelevant or uninformative words in the analysis, such as common words (stopwords).

	sentiment	content
0	NEGATIF	[malam, login, error, mulu, kontrol, daftar, o...
1	NETRAL	[daftar, online, data, input, sesuai, ktp, kk,...
2	NEGATIF	[jujur, apk, nya, nyusahin, butuhin, login, ce...
3	NEGATIF	[daftar, pakai, nomer, telfon, 8org, beda, gab...
4	NEGATIF	[update, gak, buka, , klo, seandai, nya, kondi...

Image 4. Filtering Results

Stemming

Stemming transforms words into their base form by removing affixes or

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. IX No 4, September 2023, hlm. 665-672 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v9i4.2539
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

endings, so that words with the same root can be considered as a single entity.

	sentiment	content
0	NEGATIF	malam login error mulu kontrol daftar online a...
1	NETRAL	daftar online data input sesuai ktp kk tinggal...
2	NEGATIF	jujur apk nya nyusahin butuhin login cetak kar...
3	NEGATIF	daftar pakai nomer telfon 8org beda gabisa pdh...
4	NEGATIF	update gak buka klo anda nya kondisi darurat ...

Image 5. *Stemming Results*

Term Frequency-Inverse Document Frequency (TF-IDF)

Term frequency (TF) adalah jumlah kata/term dalam suatu dokumen, sedangkan *Inverse Document Frequency* (IDF) adalah frekuensi kemunculan kata/term diseluruh dokumen[8]. TF–IDF bermanfaat dalam meningkatkan performa model, rumus TF-IDF 1.

$$\text{TF} - \text{IDF} = \frac{n_{yx}}{\sum t_y} .\log \frac{\sum d}{n_{dx}} \qquad (1)$$

x = word number x
t = term / word
y = document number y
d = document
n = number/count

Evaluasi Permodelan

In this study, the performance measurement used is the Confusion Matrix, which is a table used to evaluate the performance of a classification model or prediction algorithm. This table compares the predicted values from the model with the actual values of the observed data.

The Confusion Matrix consists of four main components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) [11]. After calculating the Confusion Matrix, performance metrics in algorithm modeling can be measured using formulas such as accuracy (2), precision (3), recall (4), and F1-score (5).

$$\text{Accuracy} = \frac{TP + TN}{(TP+FP+FN+TN)} \times 100\% \qquad (2)$$

$$\text{Precision} = \frac{TP}{TP+FP} \times 100\% \qquad (3)$$

$$\text{Recall} = \frac{TP}{TP+FN} \times 100\% \qquad (4)$$

$$\text{F1- Score} = \frac{2 \times precision \times recall}{precision + recall} \times 100\% \qquad (5)$$

RESULT AND DISCUSSION

Splitting data

The preprocessed data is then split into training data and test data. To determine the optimal result, the data is divided into 5 experiments: in experiment 1, the test data is 10% and the training data is 90%; in experiment 2, the test data is 20% and the training data is 80%; in experiment 3, the test data is 30% and the training data is 70%; in experiment 4, the test data is 40% and the training data is 60%; and finally, in experiment 5, the training data is 50% and the test data is 50%.

Naïve Bayes Classifier

In each experiment, the division results in different accuracies, as shown in Table 1.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. IX No 4, September 2023, hlm. 665-672 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v9i4.2539
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

Table 1. Naïve Bayes Results

Experiment	Results of Naive Bayes Classification Experiment			
	Percision	Recall	F1-Score	Accuracy
1	88%	87%	86%	87%
2	86%	85%	85%	86%
3	86%	85%	85%	86%
4	86%	85%	85%	85%
5	83%	82%	81%	81%

From the table data 1, the most optimal result is obtained in the first experiment with an accuracy of 87%, precision of 88%, recall of 87%, and F1-score of 86%. The implementation of the most accurate classification resulted in the following Confusion Matrix.

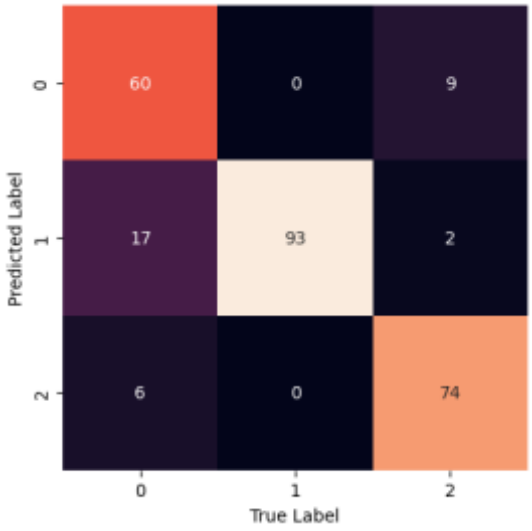

Image 6. *Confusion Matrix Naïve Bayes*

In Image 6, the number of data with a positive value and correctly predicted as positive is 60, the number of data with a neutral value and correctly predicted as neutral is 93, the number of data with a negative value and correctly predicted as negative is 79. The number of positive data predicted as neutral is 17, the number of positive data predicted as negative is 6, the number of negative data predicted as positive is 2, and the number of negative data predicted as neutral is 4.

```
MultinomialNB Accuracy: 0.8697318007662835
MultinomialNB Precision: 0.8749741200828157
MultinomialNB Recall: 0.8644932671863926
MultinomialNB f1_score: 0.864586818116985
confusion matrix:
 [[60 17  6]
 [ 0 93  0]
 [ 9  2 74]]
================================================

              precision    recall  f1-score   support

     NEGATIF       0.87      0.72      0.79        83
     NETRAL        0.83      1.00      0.91        93
     POSITIF       0.93      0.87      0.90        85

    accuracy                           0.87       261
   macro avg       0.87      0.86      0.86       261
weighted avg       0.87      0.87      0.87       261
```

Image 7. Naïve Bayes Classification Results

Support Vector Machine (SVM)

In each experiment, the division results in different accuracies, as shown in Table 2.

Table 2. Support Vector Machine Results

Experiment	Results of Support Vector Machine Experiment			
	Percision	Recall	F1-Score	Accuracy
1	97%	88%	92%	95%
2	95%	95%	94%	94%
3	94%	85%	89%	93%
4	90%	86%	88%	92%
5	90%	85%	88%	92%

Based on the table data 2, the most optimal result is obtained in the first experiment with an accuracy of 95%, precision of 97%, recall of 88%, and F1-score of 92%. The implementation of the most accurate classification resulted in the following Confusion Matrix.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. IX No 4, September 2023, hlm. 665-672 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v9i4.2539
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

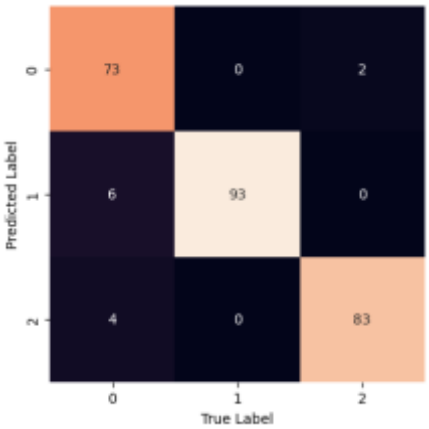

Image 8. *Confusion Matrix* SVM

In Image 8, the number of positive data correctly predicted as positive is 74, the number of neutral data correctly predicted as neutral is 93, the number of negative data correctly predicted as negative is 85. The number of positive data predicted as neutral and correctly classified is 6, and the number of positive data predicted as negative and correctly classified is 3.

```
Support Vector Machine Accuracy: 0.9540229885057471
Support Vector Machine Precision: 0.9555834204110066
Support Vector Machine Recall: 0.9519962201748169
Support Vector Machine F1 score: 0.9526389706603866
Confusion matrix:
[[73  6  4]
 [ 0 93  0]
 [ 2  0 83]]
================================================

              precision    recall  f1-score   support

     NEGATIF       0.97      0.88      0.92        83
      NETRAL       0.94      1.00      0.97        93
     POSITIF       0.95      0.98      0.97        85

    accuracy                           0.95       261
   macro avg       0.96      0.95      0.95       261
weighted avg       0.95      0.95      0.95       261
```

Image 9. SVM Classification Results

K-Nearest Neighbor (K-NN)

In each experiment, the division results in different accuracies, as shown in Table 3.

Table 3. K-Nearest Neighbor Results

Experiment	Results of *K-Nearest Neighbor* Experiment			
	Precision	Recall	F1-Score	Accuracy
1	83%	73%	68%	75%
2	82%	75%	67%	73%
3	81%	72%	66%	72%
4	81%	71%	66%	72%
5	80%	70%	64%	70%

From the table data 3, the most optimal result is obtained in the first experiment with an accuracy of 75%, precision of 83%, recall of 73%, and F1-score of 68%. The implementation of the most accurate classification resulted in the following Confusion Matrix..

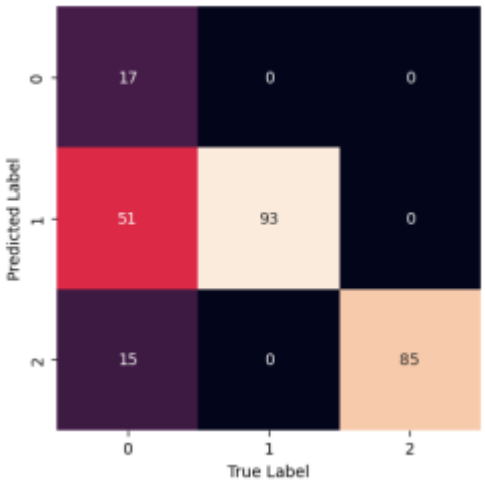

Image 10. *Confusion Matrix* KNN

In Image 10, the number of positive data correctly predicted as positive is 19, the number of neutral data correctly predicted as neutral is 93, the number of negative data correctly predicted as negative is 85. The number of positive data predicted as neutral and correctly classified is 48, and the number of positive data predicted as negative and correctly classified is 16.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. IX No 4, September 2023, hlm. 665-672 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v9i4.2539
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

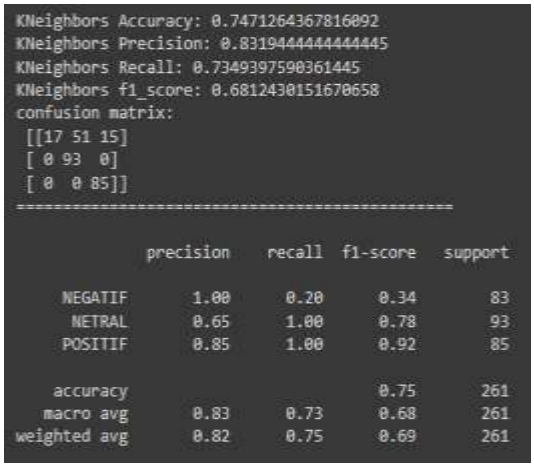

Image 11. KNN Clasification Results.

Overall Classification Results

From the five experiments conducted with three algorithms, the best result is obtained in the experiment with 10% test data and 90% training data. The overall classification results include precision, recall, F1-score, and accuracy.

Table 4. Overall Classification Results

algorithms	Overall Classification Results			
	Percision	Recall	F1-Score	Accuracy
NB	88%	87%	86%	87%
SVM	97%	88%	92%	95%
KNN	83%	73%	68%	75%

From the comparison table 4, it can be observed that the SVM classification yields higher values compared to Naive Bayes and KNN. The SVM results show an accuracy of 95%, precision of 96%, recall of 95%, and F1-score of 95%. On the other hand, Naive Bayes achieves an accuracy of 87%, precision of 88%, recall of 87%, and F1-score of 86%, while KNN achieves an accuracy of 75%, precision of 83%, recall of 73%, and F1-score of 68%.

CONCLUSION

Based on the testing results of sentiment analysis on the JKN Mobile application using three algorithms, it can be concluded that out of the 1200 SMOTE-processed datasets with different proportions of training and test data, the experiment with 10% test data and 90% training data yields the most optimal results. The accuracies for Naive Bayes, Support Vector Machine (SVM), and K-Nearest Neighbor (KNN) in this experiment are 87%, 95%, and 75%, respectively. In the context of sentiment analysis for the JKN Mobile application, it is recommended to use the Support Vector Machine (SVM) algorithm due to its higher accuracy compared to Naive Bayes and K-Nearest Neighbor. Utilizing the Support Vector Machine can assist in predicting sentiments more accurately.

BIBLIOGRAPHY

[1] Humas, “BPJS Kesehatan Ikuti Perkembangan Zaman, Mobile JKN Satu Genggaman Untuk Berbagai Kemudahan,” 2020. https://www.bpjs-kesehatan.go.id/bpjs/post/read/2020/1671/Ikuti-Perkembangan-Zaman-Mobile-JKN-Satu-Genggaman-Untuk-Berbagai-Kemudahan (accessed Jul. 02, 2023).

[2] B. Kesehatan, “Mobile JKN,” *Google Play Store*, 2023. https://play.google.com/store/apps/details?id=app.bpjs.mobile (accessed Mar. 24, 2023).

[3] J. P. Tanjung, F. C. Tampubolon, A. W. Panggabean, and M. A. A. Nandrawan, “Customer

Classification Using Naive Bayes Classifier With Genetic Algorithm Feature Selection," *Sinkron*, vol. 8, no. 1, pp. 584–589, Feb. 2023, doi: 10.33395/sinkron.v8i1.12182.

[4] N. H. Ovirianti, M. Zarlis, and H. Mawengkang, "Support Vector Machine Using A Classification Algorithm," *SinkrOn*, vol. 7, no. 3, pp. 2103–2107, 2022, doi: 10.33395/sinkron.v7i3.11597.

[5] D. Cheng, S. Zhang, Z. Deng, Y. Zhu, and M. Zong, "κ NN algorithm with data-driven k value," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8933, pp. 499–512, 2014, doi: 10.1007/978-3-319-14717-8_39.

[6] A. Nugroho and E. Rilvani, "Penerapan Metode Oversampling SMOTE Pada Algoritma Random Forest Untuk Prediksi Kebangkrutan Perusahaan," *Techno.Com*, vol. 22, no. 1, pp. 207–214, 2023, doi: 10.33633/tc.v22i1.7527.

[7] S. Lestari, M. Mupaat, and A. Erfina, "Analisis Sentimen Masyarakat Indonesia terhadap Pemindahan Ibu Kota Negara Indonesia pada Twitter," *JUSIFO (Jurnal Sist. Informasi)*, vol. 8, no. 1, pp. 13–22, 2022, doi: 10.19109/jusifo.v8i1.12116.

[8] R. Puspitasari, Y. Findawati, M. A. Rosid, P. S. Informatika, and U. M. Sidoarjo, "SENTIMENT ANALYSIS OF POST-COVID-19 INFLATION BASED ON TWITTER USING THE K-NEAREST NEIGHBOR AND SUPPORT VECTOR MACHINE ANALISIS SENTIMEN TERHADAP INFLASI PASCA COVID-19 BERDASARKAN TWITTER DENGAN METODE KLASIFIKASI K-NEAREST NEIGHBOR DAN," vol. 4, no. 4, pp. 1–11, 2023, doi: 10.20884/jutif.

[9] E. Indrayuni, A. Nurhadi, and D. A. Kristiyanti, "Implementasi Algoritma Naive Bayes, Support Vector Machine, dan K-Nearest Neighbors untuk Analisa Sentimen Aplikasi Halodoc," *Fakt. Exacta*, vol. 14, no. 2, p. 64, 2021, doi: 10.30998/faktorexacta.v14i2.9697.

[10] Z. Zaenal and I. R. I. Astutik, "Sentiment Analysis of OYO App Reviews Using the Support Vector Machine Algorithm," *Procedia Eng. Life Sci.*, vol. 3, no. December, 2023, doi: 10.21070/pels.v3i0.1338.

[11] D. S. Putri, A. Sentimen, U. Aplikasi, and T. Ridwan, "Analisis Sentimen Ulasan Aplikasi Pospay dengan Algoritma Support Vector Machine," no. 2018, 2023.