JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. X No 3, Juni 2024, hlm. 583 – 590 ISSN 2550- 020 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v10i3.3201
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

SENTIMENT ANALYSIS OF PEGIPEGI.COM ON GOOGLE PLAYSTORE WITH NAÏVE BAYES ALGORITHM

Riski Hardian[1*], Luzi Dwi Oktaviana[2], Aulia Hamdi[1]
[1]Informatic, Universitas Amikom Purwokerto
[2]Information System, Universitas Amikom Purwokerto
email: *me.riski.hardian@gmail.com

Abstract: Today, many users use online platforms rather than offline platforms for ticket bookings, involving a wide range of services such as flights, hotels, trains, buses, and entertainment. PegiPegi.com, as one of the fastest growing online travel agencies in Indonesia, demonstrates success by understanding the value of technology and maintaining strong partnerships. Users of this platform often provide reviews, viewing user reviews can be done manually but this will have a less effective impact, so it needs to be done automatically with sentiment analysis. This research the Naïve Bayes method in sentiment analysis of PegiPegi.com reviews, with a focus on understanding customer satisfaction and service improvement. By combining these approaches, this research contributes to a deeper understanding of user responses to OTA services and presents the evaluation results of the Multinomial Naive Bayes classification model with an accuracy rate of 89.5%. The high precision in the Negative class demonstrates the model's ability to identify negative reviews. However, there are challenges in classifying the Neutral class, indicating the potential for further improvement. Nevertheless, the F1 score of 0.522 reflects a good balance between overall precision, recall so it can be concluded the naïve bayes algorithm is successful for performing sentiment analysis.

Keywords: Sentiment analysis; naïve bayes algorithm; pegipegi.com; playstore

Abstract: Saat ini banyak pengguna platform online dibandingkan offline untuk pemesanan tiket, yang melibatkan berbagai layanan seperti penerbangan, hotel, kereta api, bus, dan hiburan. PegiPegi.com, sebagai salah satu agen perjalanan online yang berkembang pesat di Indonesia, menunjukkan keberhasilan dengan memahami nilai teknologi dan mempertahankan kemitraan yang kuat. Pengguna platform ini sering memberikan ulasan, melihat ulasan pengguna bisa saja dilakukan secara manual tetapi hal ini akan memberikan dampak yang kurang efektif, sehingga perlu dilakukan secara otomatis dengan analisis sentiment. Penelitian ini bertujuan untuk menerapkan metode klasifikasi Naïve Bayes dalam analisis sentimen ulasan PegiPegi.com, dengan fokus pada pemahaman kepuasan pelanggan dan peningkatan layanan. Dengan menggabungkan pendekatan ini, penelitian ini berkontribusi pada pemahaman yang lebih dalam tentang tanggapan pengguna terhadap layanan OTA dan menyajikan hasil evaluasi model klasifikasi Multinomial Naive Bayes dengan tingkat akurasi 89,5%. Presisi tinggi di kelas Negatif menunjukkan kemampuan model untuk mengidentifikasi ulasan negatif. Namun, ada tantangan dalam mengklasifikasikan kelas Netral, menunjukkan potensi untuk perbaikan lebih lanjut. Namun demikian, skor F1 0,522 mencerminkan keseimbangan yang baik antara presisi keseluruhan dan daya ingat sehingga dapat disimpulkan algoritma naïve bayes berhasil untuk melakukan analisis sentimen.

Keywords: Analisis sentimen; naïve bayes; pegipegi.com; playstore

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. X No 3, Juni 2024, hlm. 583 – 590 ISSN 2550- 020 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v10i3.3201
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

INTRODUCTION

In this era, we are witnessing a major shift in people's behavior towards the convenience of booking tickets online through various booking platforms. This phenomenon is not limited to airline tickets, but also involves reservations for other services such as hotels, trains, buses and entertainment. One vivid example of an online platform that fulfills this need is PegiPegi.com. The company known as PT. Go Online Destination or more familiarly known as Pegipegi is an online travel agent (OTA) that is experiencing rapid development in Indonesia. PegiPegi.com are apps and websites that provide booking services for different types of travel and entertainment. Through this platform, users can easily search, compare and book tickets for hotels, flights, trains, buses and various entertainment activities [1].

Pegipegi takes effective and innovative strategic measures to ensure customer comfort and satisfaction on their journeys. In order for pegipegi services to be better, an in-depth analysis process is needed by looking at both positive and negative complaints from users, seeing comments can be done manually but this is less effective, one way can be done with sentiment analysis [2].

Sentiment analysis is the process of extracting and assessing sentiment or opinion from text, usually in positive, negative, or neutral form. For this research, it will use the Naïve Bayes algorithm because of its superiority in classifying texts. The Naïve Bayes algorithm has a high computational speed, as well as being easy to implement [3]. Thus, the Naïve Bayes algorithm is the right choice to perform sentiment analysis on text data such as the PengiPegi.com review in this study.

Some previous studies were similar, the first research was research was conducted to understand people's views on paylater. Sentiment analysis using the Naive Bayes Classifier and TextBlob methods on Twitter data shows that the majority of people have a negative view of paylater. Naive Bayes Classifier is more accurate (91%) than TextBlob (61%) [4].

The two research conducted a PKKM policy analysis with the results showing an accuracy of 81.07%. Validation was performed with Support Vector Machine which resulted in 79.96% accuracy [5].

The three research conducted a covid 19 sentiment analysis Naïve Bayes had an accuracy of 85%, while K-Nearest Neighbor had an accuracy of 82% with values of k = 6, 8, and 14 [6].

The four finals of sentiment analysis research with SVM, Naive Bayes, and KNN algorithms. The evaluation showed SVM had the highest accuracy (90.01%), followed by Naive Bayes (79.20%), and KNN (62.10%) [7].

The five research of Twitter's sentiment analysis got positive results, with 60,581 positive tweets, 58,998 negative tweets, and 32,291 neutral tweets. This analysis provides insight into the public's view of the postponement of the election [8].

The six research, conducted sentiment analysis on gojek used to check user reviews. This study compares SVM and KNN methods to classify Gojek reviews on Google Playstore. The results show SVM is better in the sentiment analysis of the Gojek application [9].

The seventh research, film review research is the audience's view of a film.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. X No 3, Juni 2024, hlm. 583 – 590 ISSN 2550- 020 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v10i3.3201
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

To analyze audience response, sentiment analysis using the Naïve Bayes method was used on IMDB reviews. The results were divided into positive and negative responses, and tested using Chi Square [10].

The eight research, this study analyzed public opinion about online learning in Indonesia in November 2020 through Twitter. As a result, 30% of sentiment was positive, 69% negative, and 1% neutral. Negative sentiment is caused by dissatisfaction, mainly due to feeling 'stressed' and 'lazy' [11].

The Ninth research, Analysis of Indonesian people's sentiment towards the pre-employment card program through Twitter using the SVM method. Evaluation shows that public sentiment tends to be neutral (98.34%), with linear kernels having higher accuracy (98.67%) than RBF kernels [12].

The tenth research Sentiment analysis of tweets and comments on social media. Of the 1075 data, 28% were positive, 27% were negative, and 45% were neutral. In conclusion, public sentiment tends to be balanced between positive and negative, with the majority neutral [13]

This research is very different from some of the research that has been described previously, later this research will obtain the results of negative and positive reviews of PegiPegi.com users, this makes it easier as an application evaluation material while previous research only focuses on one of them, for example negative or positive reviews, besides that the accuracy value of the application of the naïve bayes algorithm and the matrix obtained. Not only the accuracy value of this research is complemented by the acquisition of recall value, f1 value and precision value. This research is successful and very complete so it can be concluded that there are many elements of novelty and different from other studies.

METHOD

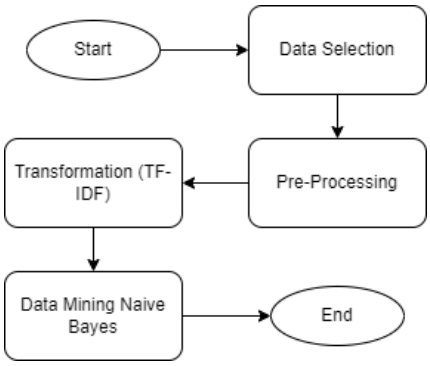


Image 1. research steps

Image 1. illustrates the process of analyzing text data using the Naive Bayes algorithm. The process begins with selecting relevant data, including data sources and required variables [14].

Data Selection

This initial stage involves information gathering and statistical labeling. Data is retrieved from the Pegipegi application on the Google Play Store using web scraping techniques with the Python programming language [15].

Data Pre-Processing

This stage is the first step in sentiment analysis, where unstructured data is converted into structured data before the main analysis is performed [16].

Cleaning

Cleaning was done to remove emoticons and symbols from reviews because the focus of this study was more on text. Deleted characters include "~", " '", "!", " $", " and more. For example, a

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. X No 3, Juni 2024, hlm. 583 – 590 ISSN 2550- 020 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v10i3.3201
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

review like "min help, fix delivery service" would be "min please for delivery service improvement".

Case Folding

Case folding converts all letters in the text to lowercase. For example, "Please upgrade payment features" will be changed to "please upgrade payment features".

Word Normalization

Word normalization improves the wording in the review to conform to the correct Indonesian system. For example, "why does the update process take so long" is changed to "why does the update process take so long".

Stopword Removal

Stopword deletion is the stage of eliminating words based on a list of conjunctions. Words like “in”, “and”, “the” will be removed.

Tokenization

Tokenization is a technique used to break text into words, taking into account punctuation and spacing boundaries. For example, the sentence "shopee is slow now" was then changed to "shopee", "now", "slow".

Stemming

Stemming is the process of reducing words to their base form or “root” word. For example, the word "cancel" would be changed to "cancel". Stemming is used in text processing and text analysis to help reduce word variation and increase consistency.

Transformation

The transformation or attribute formation stage in text refers to the process of obtaining the required representation. At this stage, the author carried out feature extraction using the TF-IDF method. The following is the formula for TF – IDF.

$$y: W(d,t) = TF(d,t) \qquad (1)$$

In this equation it can be explained that TF (d, t) is the frequency of term t in text d.

Data Mining

In the field of computer science, data mining refers to data mining techniques that aim to discover hidden patterns in a data set with the aim of generating new knowledge. Specifically, data mining involves various methods adapted to the intended use of the data, such as estimation, prediction, classification, clustering, and association [17]

RESULT AND DISCUSSION

Data Selection

This step is part of the preparation in the data selection process. Information obtained from interviews will undergo an attribute selection stage before then undergoing a preprocessing process.

Data Scrapping

Scraping process is included in the data mining stage category, which involves the integration of various scientific disciplines such as machine learning, pattern recognition, statistics, databases, and visualization. Data mining aims to uncover hidden information in data collections or datasets.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. X No 3, Juni 2024, hlm. 583 – 590 ISSN 2550- 020 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v10i3.3201
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

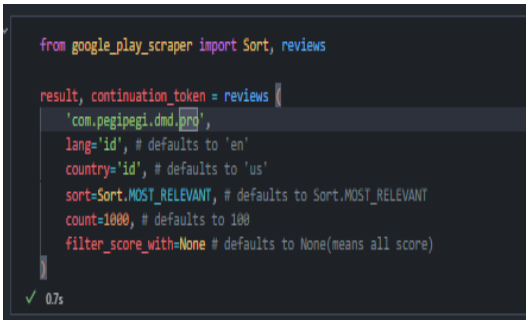


Image 2. Data scrapping process

Image 2 utilizes the google_play_scraper` library to access reviews of the Pegipegi application in the Google Play Store. Through the `reviews` function, applications are identified based on the package ID 'com.pegipegi.dmd.pro', with the review language set to Indonesian ('id') and the country set to Indonesia ('id'). Reviews are sorted by relevance using the `Sort.MOST_RELEVANT` parameter, and 1000 reviews are taken with the `count=1000` parameter. No review score filter is applied (`None`), so all review scores are taken. The result of this function includes two values, namely `result` which contains a list of reviews and related information, and `continuation_token` which can be used to continue fetching reviews if necessary. With this configuration, we can collect user reviews of the Pegipegi application from the Google Play Store for further analysis.

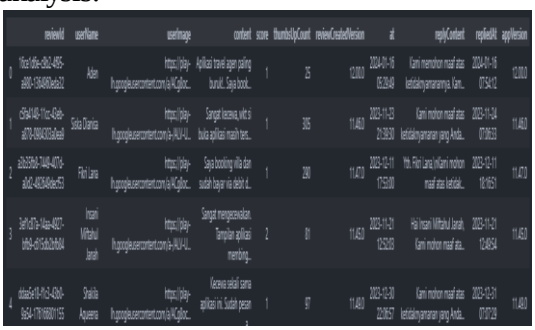
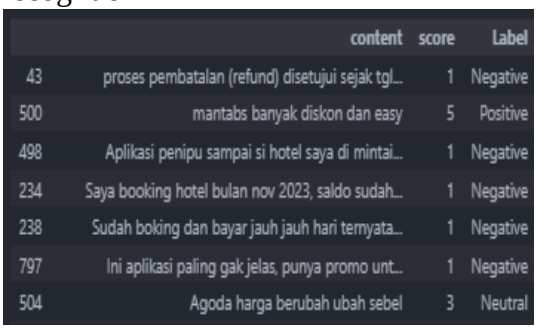


Image 3. results from scrapping data

Data from Image 3 can be used by developers or researchers to gain further insight from user reviews of the Pegipegi application on the Google Play Store. This information can serve as a basis for evaluating application performance, planning feature updates, or improving services based on feedback received from users.

Labeling data

The data labeling process is critical in a variety of machine learning use cases, including computer vision, natural language processing, and speech recognition



	content	score	Label
43	proses pembatalan (refund) disetujui sejak tgl...	1	Negative
500	mantabs banyak diskon dan easy	5	Positive
498	Aplikasi penipu sampai si hotel saya di mintai...	1	Negative
234	Saya booking hotel bulan nov 2023, saldo sudah...	1	Negative
238	Sudah boking dan bayar jauh jauh hari ternyata...	1	Negative
797	Ini aplikasi paling gak jelas, punya promo unt...	1	Negative
504	Agoda harga berubah ubah sebel	3	Neutral

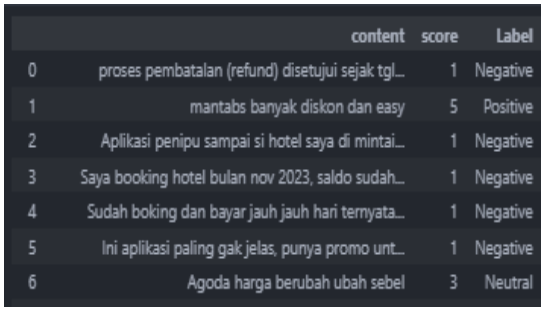
Image 4. Results from Data Labeling

Image 4 shows the results of the data labeling process. The results of data labeling are very instrumental in developing a sentiment analysis model that is accurate and appropriate to the specific context of Pegipegi.com.

Data Pre-Processing

The pre-processing process in the Naive Bayes algorithm includes a series of steps involving data cleaning by handling missing values, tokenization to separate word units, converting letters to lowercase, eliminating common words, and applying a stemming or lemmatization process.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. X No 3, Juni 2024, hlm. 583 – 590 ISSN 2550- 020 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v10i3.3201
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi



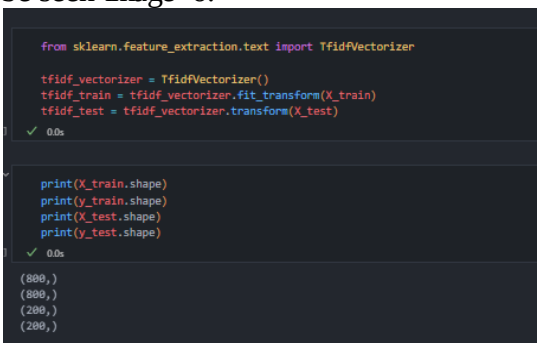
	content	score	Label
0	proses pembatalan (refund) disetujui sejak tgl...	1	Negative
1	mantabs banyak diskon dan easy	5	Positive
2	Aplikasi penipu sampai si hotel saya di mintai...	1	Negative
3	Saya booking hotel bulan nov 2023, saldo sudah...	1	Negative
4	Sudah boking dan bayar jauh jauh hari ternyata...	1	Negative
5	Ini aplikasi paling gak jelas, punya promo unt...	1	Negative
6	Agoda harga berubah ubah sebel	3	Neutral

Image 5. shows the output of the pre-processing process

Where the text has been processed to become more structured, standardized, and ready to be used in the next analysis stage. Data that has undergone pre-processing has a higher potential to produce accurate and relevant analysis, especially when applied in machine learning models such as the Naive Bayes sentiment classification algorithm.

Transformation

TF-IDF weighting and results can be seen image 6.



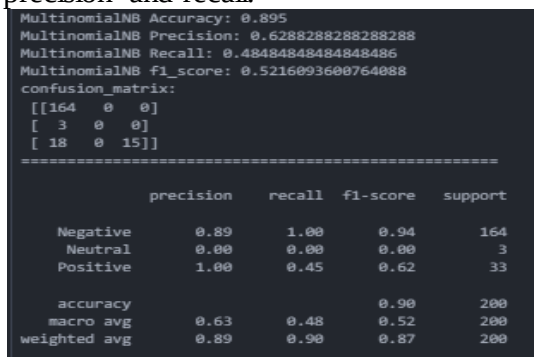
```
from sklearn.feature_extraction.text import TfidfVectorizer

tfidf_vectorizer = TfidfVectorizer()
tfidf_train = tfidf_vectorizer.fit_transform(X_train)
tfidf_test = tfidf_vectorizer.transform(X_test)
✓ 0.0s

print(X_train.shape)
print(y_train.shape)
print(X_test.shape)
print(y_test.shape)
✓ 0.0s
(800,)
(800,)
(200,)
(200,)
```

Image 6. TF-IDF weighting and results

Image 6 shows the weighting process using the Term Frequency-Inverse Document Frequency (TF-IDF) method using the `scikit-learn` library. In the code, the `TfidfVectorizer` object from `scikit-learn` is used to convert text to a numerical representation based on the frequency of words in the document and the inverse document frequency of those words. The training data (`X_train`) and test data (`X_test`) are then converted into a TF-IDF matrix using the `fit_transform` and `transform` methods in the `TfidfVectorizer` object. The results of this transformation produce a numerical representation that can be used as input for training and testing machine learning models, such as the Naive Bayes classification algorithm. This TF-IDF representation can be used in a variety of text processing tasks, including sentiment analysis.

Data Mining

The evaluation results of the Multinomial Naive Bayes classification model show generally good performance, with an accuracy rate of 89.5%. The high precision in the Negative class (164 out of 164) indicates that the model is very good at identifying truly negative reviews. However, it should be noted that the model tends to face difficulties in classifying the Neutral class, with the precision and recall of this class each reaching 0.00. The overall F1-score of 0.522 reflects a good balance between precision and recall.



```
MultinomialNB Accuracy: 0.895
MultinomialNB Precision: 0.6288288288288288
MultinomialNB Recall: 0.48484848484848486
MultinomialNB f1_score: 0.5216093600764088
confusion_matrix:
 [[164   0   0]
 [  3   0   0]
 [ 18   0  15]]
==================================================

              precision    recall  f1-score   support

    Negative       0.89      1.00      0.94       164
     Neutral       0.00      0.00      0.00         3
    Positive       1.00      0.45      0.62        33

    accuracy                           0.90       200
   macro avg       0.63      0.48      0.52       200
weighted avg       0.89      0.90      0.87       200
```

Image 7. research results

Image 7 is the result of research that has been carried out. The confusion matrix presents a comparison between the actual classification results and the predicted results. From the confusion matrix, it can be seen that the model tends to focus more on the Negative

class, but has challenges in recognizing the Neutral class. This model evaluation provides valuable insight into model performance and can be used as a foundation for further improvements, particularly in addressing class imbalance and improving the ability to recognize reviews with neutral sentiment.

CONCLUSION

The conclusion of the study includes an in-depth understanding of the trend of consumers switching to online ticket reservations through platforms such as PegiPegi.com. Using the Naïve Bayes classification method, this study aims to contribute to improving service quality, maintaining consistency, and increasing user trust. Furthermore, the evaluation results of the Multinomial Naive Bayes classification model showed generally good performance, with an accuracy rate of 89.5%. The high precision of the Negative class (164 out of 164) indicates that the model is very good at identifying truly negative reviews. However, it should be noted that these models tend to face difficulties in classifying Neutral classes, with precision and recall reaching 0.00 each. An overall F1 score of 0.522 reflects a good balance between precision and recall, indicating the model's ability to deliver balanced performance in identifying positive and negative reviews.

BIBLIOGRAPHY

[1] D. R. Alghifari, M. Edi, And L. Firmansyah, "Implementasi Bidirectional Lstm Untuk Analisis Sentimen Terhadap Layanan Grab Indonesia," *Jurnal Manajemen Informatika (Jamika)*, Vol. 12, No. 2, Pp. 89–99, Sep. 2022, Doi: 10.34010/Jamika.V12i2.7764.

[2] D. Darwis, E. Shintya Pratiwi, A. Ferico, And O. Pasaribu, "Penerapan Algoritma Svm Untuk Analisis Sentimen Pada Data Twitter Komisi Pemberantasan Korupsi Republik Indonesia," 2020.

[3] F. Fathonah And A. Herliana, "Penerapan Text Mining Analisis Sentimen Mengenai Vaksin Covid - 19 Menggunakan Metode Naïve Bayes," *Jurnal Sains Dan Informatika*, Vol. 7, No. 2, Pp. 155–164, Dec. 2021, Doi: 10.34128/Jsi.V7i2.331.

[4] A. Safira, A. S. Masyarakat…ν, And F. N. Hasan, "Analisis Sentimen Masyarakat Terhadap Paylater Menggunakan Metode Naive Bayes Classifier," *Jurnal Sistem Informasi*, Vol. 5, No. 1, 2023.

[5] C. F. Hasri And D. Alita, "Penerapan Metode Naïve Bayes Classifier Dan Support Vector Machine Pada Analisis Sentimen Terhadap Dampak Virus Corona Di Twitter," *Jurnal Informatika Dan Rekayasa Perangkat Lunak (Jatika)*, Vol. 3, No. 2, Pp. 145–160, 2022, [Online]. Available: Http://Jim.Teknokrat.Ac.Id/Index. Php/Informatika

[6] K. Keahlian, R. Data, A. Luthfika Fairuz, R. Dias Ramadhani, N. Annisa, And F. Tanjung, "Jurnal Dinda Analisis Sentimen Masyarakat Terhadap Covid-19 Pada Media Sosial Twitter," 2021. [Online]. Available: Http://Journal.Ittelkom-Pwt.Ac.Id/Index.Php/Dinda

[7] F. S. Pamungkas And I. Kharisudin, "Analisis Sentimen Dengan Svm," Vol. 4, Pp. 628–634, 2021, [Online]. Available: Https://Journal.Unnes.Ac.Id/Sju/Index.Php/Prisma/

[8] A. Perdana, A. Hermawan, And D. Avianto, "Analisis Sentimen Terhadap Isu Penundaan Pemilu Di Twitter Menggunakan Naive Bayes Clasifier," *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, Vol. 11, No. 2, Pp. 195–200, Jul. 2022, Doi: 10.32736/Sisfokom.V11i2.1412.

[9] M. N. Muttaqin And I. Kharisudin, "Analisis Sentimen Pada Ulasan Aplikasi Gojek Menggunakan Metode Support Vector Machine Dan K Nearest Neighbor," *Unnes Journal Of Mathematics*, Vol. 10, No. 2, Pp. 22–27, 2021, [Online]. Available: Http://Journal.Unnes.Ac.Id/Sju/Index.Php/Ujm

[10] A. Z. Amrullah, A. Sofyan Anas, M. Adrian, And J. Hidayat, "Analisis Sentimen Movie Review Menggunakan Naive Bayes Classifier Dengan Seleksi Fitur Chi Square," *Jurnal*, Vol. 2, No. 1, 2020, Doi: 10.30812/Bite.V2i1.804.

[11] W. A. Prabowo And C. Wiguna, "Sistem Informasi Umkm Bengkel Berbasis Web Menggunakan Metode Scrum," *Jurnal Media Informatika Budidarma*, Vol. 5, No. 1, P. 149, Jan. 2021, Doi: 10.30865/Mib.V5i1.2604.

[12] N. Hendrastuty, A. Rahman Isnain, And A. Yanti Rahmadhani, "Analisis Sentimen Masyarakat Terhadap Program Kartu Prakerja Pada Twitter Dengan Metode Support Vector Machine," Vol. 6, No. 3, 2021, [Online]. Available: Http://Situs.Com

[13] H. Tuhuteru And U. Kristen Indonesia Maluku Jl Ot Pattimaipauw, "Analisis Sentimen Masyarakat Terhadap Pembatasan Sosial Berksala Besar Menggunakan Algoritma Support Vector Machine."

[14] K. M. Elistiana, Bagus Adhi Kusuma, P. Subarkah, And H. A. Awal Rozaq, "Improvement Of Naive Bayes Algorithm In Sentiment Analysis Of Shopee Application Reviews On Google Play Store," *Jurnal Teknik Informatika (Jutif)*, Vol. 4, No. 6, Pp. 1431–1436, Dec. 2023, Doi: 10.52436/1.Jutif.2023.4.6.1486.

[15] S. Lestari And S. Saepudin, "Analisis Sentimen Vaksin Sinovac Pada Twitter Menggunakan Algoritma Naive Bayes," 2021. [Online]. Available: Https://Vaksin.Kemkes.Go.Id/

[16] B. Z. Ramadhan, I. Riza, And I. Maulana, "Analisis Sentimen Ulasan Pada Aplikasi E-Commerce Dengan Menggunakan Algoritma Naïve Bayes," 2022. [Online]. Available: Http://Jurnal.Polibatam.Ac.Id/Index.Php/Jaic

[17] D. S. Utami And A. Erfina, "Analisis Sentimen Pinjaman Online Di Twitter Menggunakan Algoritma Support Vector Machine (Svm)," 2021.