Implementation of Support Vector Machine algorithm for Shopee Customer Sentiment Analysis

Melda Betaria Sitepu[1]*, Ibnu Rasyid Munthe[2], Syaiful Zuhri Harahap[3]
[1,2,3]Universitas Labuhanbatu, Indonesia
[1]meldasitepu64@gmail.com, [2]ibnurasyidmunthe@gmail.com, [3]syaifulzuhriharahap@gmail.com

Abstract:. As the number one largest marketplace in Indonesia based on the criteria for the origin of international stores, Shopee must always improve the quality of its products and services based on reviews from users. Given the huge number of user reviews, it is not effective to identify them by reading one by one. For this reason, an automated system is needed that can read and identify reviews better. Sentiment analysis has proven to do the job. This study aims to conduct a sentiment analysis of shopee product reviews from users who use English. This study applies the Support Vector Machine algorithm to classify the Shopee user review data. To solve this problem, the research was carried out by going through several stages, namely: pre-processing the text of the dataset, performing feature extraction, after that the word weighting was carried out using the TF-IDF method, after clean data was obtained, the SVM algorithm was implemented, for further evaluation of the model. In the results of the study, it was found that the word that most represented the positive opinion of Shopee customers was "Good" with a total of 4684 words. While the word that represents the most negative opinion is "Seller" with 68 words. From the five sentiment analysis models tested, the average value of the confusion matrix is obtained, which are precision=1, recall=0.97, and f1-score=0.98. From this research, it can be concluded that the SVM algorithm is proven to be applicable in conducting sentiment analysis on user reviews of Shopee products with an average accuracy rate of 97.3%.

Keywords: Classification; Sentiment Analysis; Shopee; SVM; Opinion Mining.

INTRODUCTION

The increasing development of Information and Communication Technology has affected the process of buying and selling transactions, which were originally conventional, now turned into modern ones. One type of buying and selling transactions in modern times is e-commerce. E-commerce is a channel for consumers to transact goods and services electronically via the internet (Widyastuti & Prastitya, 2020). Through e-commerce, goods and services are promoted to consumers. E-Commerce allows consumers to make purchases of goods anywhere and anytime.

Indonesia is an ASEAN country that ranks first in terms of product sales through e-commerce (Azzahro, Handayani, Murti, & Yudhoatmojo, 2020). One of the marketplace-based e-commerce that is often used by Indonesians is Shopee (Nurdin, Hutomi, Qamal, & Bustami, 2020). Shopee is one of the marketplaces with the most visitors in Indonesia with 71.5 million visits in 2020 (Limbong, Sembiring, & Hartomo, 2022). In fact, according to iprice, Shopee is the number 1 largest marketplace based on International Store origins with a monthly number of 138.7 million web visitors in the fourth quarter of 2021 (Iprice, 2021)..

As the largest marketplace application in Indonesia, Shopee provides product and service review features to users. The review feature is very useful to find out whether feedback from users is positive or negative (Rhohmawati, Slamet, & Pratiwi, 2019). The existence of user reviews of a product will be taken into consideration for buyers in deciding to purchase a product. With the review feature, it will also be input for the developer (Shopee) to improve the quality of their services. To identify a very large number of user reviews, it is not possible to do it manually by reading them one by one (Wiratama & Rusli, 2019). The process will take a very long time, and is not effective. For that we need a method that is able to identify reviews from users more effectively and efficiently. Sentiment analysis has been proven to be able to identify reviews from Shopee users

*name of corresponding author
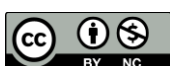
by digging in-depth information, resulting in a classification of whether the reviews are positive, negative, or neutral (Kabiru & Sari, 2019).

A number of literatures state that Sentiment Analysis is the same term as opinion mining, opinion extraction, sentiment mining, or review mining (Rolliawati, Khalid, & Rozas, 2020). Sentiment analysis is a method that can classify texts into positive, neutral, or negative opinions automatically (Saputra, Nurhadryani, Wijaya, & Defina, 2021). The purpose of sentiment analysis is to map one's opinion based on a certain topic (State, Muhardi, & Putri, 2020).

Analysis of user sentiment on product reviews and Shopee applications has been carried out by a number of previous researchers by applying various classification algorithms. The first research was conducted on Indonesian-language reviews of shopee users who bought Smartphone products with the Xiaomi Redmi Note 9 using the Naïve Bayes Classifier algorithm, the resulting accuracy rate was 85%, but there is still a lot of data that is not perfectly classified (Sihombing, Hannie, & Dermawan, 2021). Second, by applying the same algorithm, sentiment analysis was carried out on Indonesian-language reviews of Shopee application users on Playstore, resulting in an accuracy rate of 96.667% (Pratmanto et al., 2020). Third, this study applies the K-Nearest Neighbor (k-NN) algorithm with the Natural Language Processing approach to analyze the sentiments of Indonesian-speaking users towards instant hijab products at Shopee, the resulting accuracy rate is 76.92% (Muktafin, Kusrini, & Luthfi, 2020). Fourth, sentiment analysis is carried out by applying the maximum entropy classification method to Indonesian-language reviews of Shopee products on Google Play, the accuracy results obtained are 97.32% (Rhohmawati et al., 2019). Fifth, this study combines the K-means algorithm and the Naïve Bayes Classifier to conduct a sentiment analysis on user reviews of Shopee products, the accuracy results obtained are 77.12%, although the classification using K-means is considered not optimal (Hariguna, Baihaqi, & Nurwanti, 2019).

The selection of the best algorithm in conducting sentiment analysis is very important, it will affect the level of accuracy. Several studies have compared a number of classification algorithms in conducting sentiment analysis on several topics. In the analysis of sentiment comments on political figures in online media, the results show that the Support Vector Machine (SVM) algorithm has higher accuracy than the Naïve Bayes Classifier (NBC) algorithm (Kurniawan et al., 2019). Sentiment analysis of comments on Youtube on the Samsung Galaxy Z Flip 3 gadget concludes that the SVM algorithm has a better accuracy rate than NBC, which is 96.43% (Iskandar & Nataliani, 2021). Research on the sentiment analysis of MotoGP comments also places the SVM algorithm better than NBC with an accuracy rate of 95.50% (Siswanto, Wibawa, Gata, Gata, & Kusumawardhani, 2018). Other studies related to sentiment analysis also show that the SVM algorithm is better than NBC in the case of airline reviews (Rahat, Kahir, & Masum, 2019).

This research is different from several previous studies that have been described. This study aims to conduct a sentiment analysis on customer reviews of Shopee products using the SVM algorithm. The study was conducted on customer reviews in English. The dataset used as research material is secondary data sourced from Kaggle. This research will divide customer opinion into two, namely, positive opinion and negative opinion. The formulation of the problem in this study is whether the SVM algorithm can be applied in conducting sentiment analysis on Shopee customers, and how much accuracy is generated.

METHOD

This study applies the Support Vector Machine algorithm in conducting sentiment analysis on English-language reviews of Shopee product customers. The study used secondary data, namely datasets obtained from Kaggle (https://www.kaggle.com/datasets/shymammoth/shopee-reviews). The number of records from the dataset actually consists of 1,502,575 rows. For the purposes of this study, only 10,000 records were used which consisted of three attributes, namely, label, text and Sentiment. The stages of this research are shown in the research framework in Figure 1 (Hafidz & Yanti Liliana, 2021) (Somantri & Apriliani, 2018).
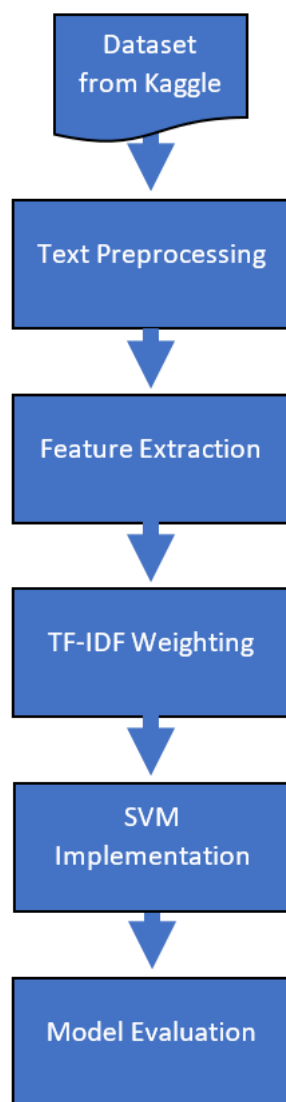
*name of corresponding author

Fig 1. Research Framework

Dataset

The dataset used in this study is an unstructured document. The first thing to do is to read it. In this dataset there are still words or sentences that have not been properly structured so that text processing still needs to be done on the dataset.

Text Preprocessing

At this stage, pre-processing is carried out to remove unnecessary data or data contained in the text that does not match the required process. The implementation of this pre-processing of text will use the Python programming language. Several stages for pre-processing text are Case Folding, Tokenization, Stopwords removal, Stemming, and Lemmatization.

Feature Extraction

At this stage, the text is converted into "bag-of words" which is used as a tool in making features. Then calculated various sizes to characterize the document. Each entry from the list refers to the frequency or count of corresponding entries in the bag-of-words list. In this phase, the CountVectorizer class is used in the Scikit-Learn Python library. Furthermore, the extracted features will be stored in a file called feature.pkl by using the picke module in the python library.

TF-IDF Weighting

*name of corresponding author

At this stage, the data is converted into a matrix and the value is weighted for each word using the Terms Frequency-Inverse Document Frequency (TF-IDF) algorithm. The results of the TF-IDF value weighting will show the similarities between the documents in the dataset.

Support Vector Machine (SVM) Implementation

The application of the SVM algorithm uses a linear kernel. The data training considers parameter C. Parameter C is used = 1. The model that has been trained with the data will then be used to classify the data.

Model Evaluation

Evaluation of model performance is carried out based on error metrics so that the accuracy of the model is obtained. To evaluate the performance of the model, the K-fold Cross-Validation (CV) method was used. The k-fold used is k=5 which consists of 5 data training groups. In this evaluation, a confusion matrix is used to obtain precision, recall, f1-score, and accuracy values. Furthermore, these values are averaged, so that the final result of this process is obtained.

RESULT

After exploring the data on the dataset, it was found that the dataset used was classified as good because there were no duplicates and nothing was null. Table 1 shows the results of exploratory data analysis in python.

Table 1. Exploratory Data Analysis

Index	Label
count	10000.0
mean	4.7646
std	0.712066
min	1.0
25%,	5.0
50%	5.0
75%	5.0
max	5.0

Table 1 shows that the dataset consists of 10,000 records with an average value of 4.76, and a standard deviation of 0.71. The standard deviation is the root of the variance, so the larger the value of the standard deviation, the more spread out the sample data (varies) from the average. The dataset used is classified as good, because there is no duplicate data and no data that is null.

Text pre-processing is done to remove unnecessary data or data contained in the text that does not match the required process. The results of text processing using Case Folding, Tokenization, Stopwords removal, Stemming, and Lemmatization methods are shown in the table below, respectively.

Table 2. Text Pre-Processing Results

Dataset	Case Folding	Tokenization	Stopwords Removal	Stemming	Lemmatization
"NOTE: DOESN'T INCLUDE THE BLUE SMALL ONE. The delivery was fast, and the quality is real good. Recommended! "	note: doesn't include the blue small one. the delivery was fast, and the quality is real good. recommended!	note doesnt include the blue small one the delivery was fast and the quality is real good recommended	note doesnt include blue small one delivery fast quality real good recommended	note doesnt include blue small one delivery fast quality real good recommend	note doesnt include blue small one delivery fast quality real good recommend
"FAST DELIVERY! Ordered on	fast delivery! ordered on 3rd jun,	fast delivery ordered on 3rd jun	fast delivery ordered 3rd jun received	fast delivery order 3rd jun receive 6th	fast delivery order rd jun receive th cant

*name of corresponding author

3rd Jun, received on the 6th. Can't wait to try it. :) Seller was really helpful and patient when I met with an issue with the courier service."	received on the 6th. can't wait to try it. :) seller was really helpful and patient when i met with an issue with the courier service.	received on the 6th cant wait to try it seller was really helpful and patient when i met with an issue with the courier service	6th cant wait try seller really helpful patient met issue courier service	cant wait try seller really helpful patient met issue courier service	wait try seller really helpful patient met issue courier service

The results of the case folding process show that there has been a change in the shape of the letters of the dataset, from using capital letters to lowercase letters. The result of the tokenization process using the remove_punctuation library in python has removed all punctuation from strings or user reviews in the dataset. The results of the stopwords removal process have reduced words that are considered unimportant in the dataset such as the words "the", "was", "and", and "is". The result of the stemming process has eliminated word affixes such as prefixes, infixes, suffixes, and prefixes and confixes in the dataset. In this test dataset the "ed" suffix has been omitted. The result of the stemming process has changed the words in the dataset back to their basic form.

The next process is implementing SVM through TF-IDF weighting on each word. In this test, a linear kernel is used with the parameter C=1. This process conducts sentiment model training to classify data. To evaluate the performance of the model, the Cross-Validation (CV) method is used in the K-fold python library. In this implementation, 5 data training groups were used (k=5). The results of the sentiment analysis are shown in Figure 3 below.

```
              precision    recall  f1-score   support

    Negative       0.03      0.67      0.05         3
    Positive       1.00      0.96      0.98      1997

    accuracy                           0.96      2000
   macro avg       0.51      0.81      0.51      2000
weighted avg       1.00      0.96      0.98      2000

model_sentimen 4  saved in ./model.pkl 4
```

Fig 2. Sentiment Analysis Results

Based on the classification results from the sentiment analysis that has been carried out, it appears that there are 5 repetitions. In model_sentiment 4 the average value (weight avg) of precision, recall, f1-score and support is stable (doesn't change too much). This shows that the classification process is good with an accuracy value of 96%. After analyzing the sentiment of the shopee user customer data. Then the data plot is carried out on the results of the classification of sentiment analysis as shown in Figure 3.
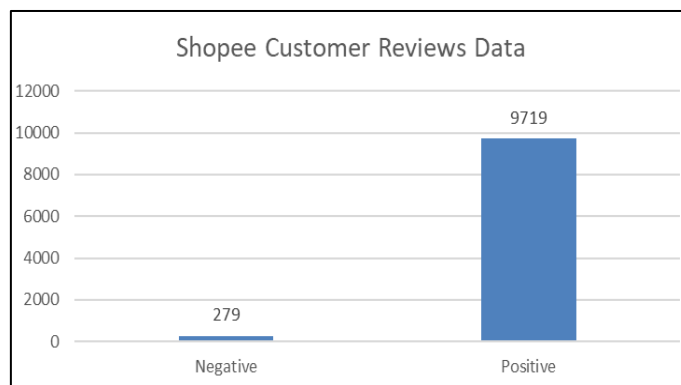
Fig 3. Customer Opinion Classification

*name of corresponding author

It can be seen in Figure 3 that the reviews from shopee customers are divided into two opinions, namely, "Positive" and "Negative" opinions. Of the 10,000 user data records that have been tested, the results show that the positive opinions of Shopee customers amount to 9719 data with a percentage of 97.19%. While the negative opinions of Shopee customers amounted to 279 data with a percentage of 27.9%.

To see the occurrence of words that represent each positive and negative opinion, a data visualization plot is carried out using the World Cloud as shown in Figures 4 and 5.

Fig 4. Positive Opinion Visualization

Fig 5. Negative Opinion Visualization

To see the number of occurrences of words in positive and negative opinions, a data plot is carried out on the 10 highest words representing each opinion. The data plots are shown in Figures 6 and 7.
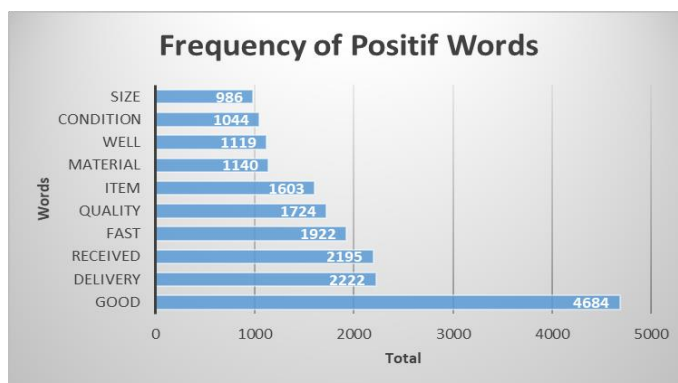

Fig 6. Frequency Positive Words

*name of corresponding author

From Figure 5 we can find out 10 words that often appear in positive opinions of Shopee product customer reviews. The word that represents the most positive opinion is "Good" with 4684 words from 10,000 datasets.
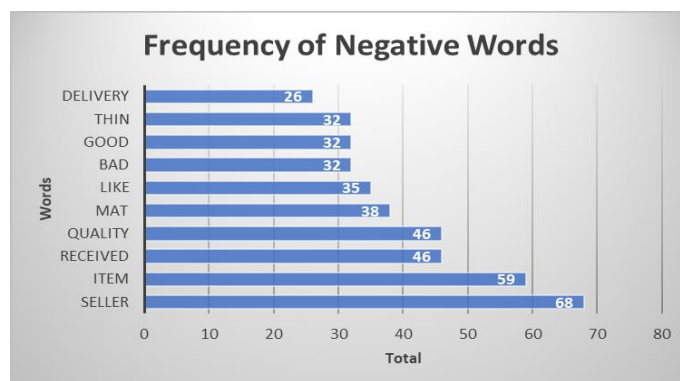

Fig 7. Frequency Negative Words

From Figure 5 we can see 10 words that often appear in negative opinions of Shopee product customer reviews. The word that represents the most negative opinion is "Seller" with 68 words from 10,000 datasets.

The next process is to evaluate the sentiment model that has been implemented. The evaluation of the model uses a confusion matrix that produces precision, recall, f1-score, and accuracy values. For more details, it is presented in Table 3 below.

Table 3. Evaluasi Model Analisis Sentimen SVM

K-Fold	Precision	Recall	F1-Score	Accuracy
k=0	1.0	0.98	0.99	0.98
k=1	1.0	0.98	0.99	0.98
k=2	1.0	0.97	0.98	0.97
k=3	1.0	0.98	0.99	0.98
k=4	1.0	0.96	0.98	0.96

From Table 3 above, it can be seen that the results for each model are presented with values of precision, recall, f1-score, and accuracy. The highest scores were obtained at k-fold 0, 1, and 3 with each value for precision is 1, recall is 0.98, f1-score is 0.99, and accuracy is 0.98. From the five sentiment analysis models, the average value was sought, the results obtained were precision = 1, recall = 0.97, f1-score = 0.98, and accuracy = 0.97.

DISCUSSIONS

This research has succeeded in building a sentiment analysis model by applying the SVM algorithm. Where done 5 repetitions on the training model with k-fold = 5. Regarding the difference in results between each k-fold, it is not yet known what is the real causal factor. Is it because of the difference in the average weight that appears or because of the weighting carried out in the TF-IDF process.

This research is still fairly simple, because it only tests 1 algorithm, namely, SVM. In the future, more comprehensive research should be carried out by comparing more than one classification algorithm and combined with optimization methods, so that sentiment analysis on Shopee product customer reviews can be more perfect.

CONCLUSION

This research has succeeded in conducting a sentiment analysis on customer reviews of Shopee products that use English. The results of the study conclude that the algorithm can be applied to perform sentiment analysis on Shopee customers. Sentiment model testing was carried out 5 times with the average precision, recall, f1-score being 1.0; 0.97; 0.98. The average level of accuracy produced is 97.3%.

The use of the Support Vector Machine algorithm with TF-IDF weighting can be one way to solve classification problems in sentiment analysis. This can be proven by the use of word weighting results using the TF-IDF method. The success of the SVM algorithm cannot be separated from text preprocessing by means of Case Folding, Tokenization, Stopwords removal, Stemming, and Lemmatization..

*name of corresponding author

REFERENCES

Azzahro, F., Handayani, P. W., Murti, S. S., & Yudhoatmojo, S. B. (2020). The Effects of Perceived Justice and Emotions on Service Recovery Satisfaction on Indonesian B2B and C2C E-commerce Customers. *Jurnal Sistem Informasi*, *16*(1), 38–48.

Hafidz, N., & Yanti Liliana, D. (2021). Klasifikasi Sentimen pada Twitter Terhadap WHO Terkait Covid-19 Menggunakan SVM, N-Gram, PSO. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(2), 213–219. https://doi.org/10.29207/resti.v5i2.2960

Hariguna, T., Baihaqi, W. M., & Nurwanti, A. (2019). Sentiment Analysis of Product Reviews as A Customer Recommendation Using the Naive Bayes Classifier Algorithm. *International Journal of Informatics and Information Systems*, *2*(2), 48–55. https://doi.org/10.47738/ijiis.v2i2.13

Iprice. (2021). Peta E-Commerce Indonesia. Retrieved April 18, 2022, from iprice.co.id website: https://iprice.co.id/insights/mapofecommerce/

Iskandar, J. W., & Nataliani, Y. (2021). Perbandingan Naïve Bayes, SVM, dan k-NN untuk Analisis Sentimen Gadget Berbasis Aspek. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(6), 1120–1126. https://doi.org/10.29207/resti.v5i6.3588

Kabiru, I. N., & Sari, P. K. (2019). Analisa Konten Media Sosial E-Commerce Pada Instagram Menggunakan Metode Sentimen Analysis Dan LDA-Based Topic Modeling (Studi Kasus : Shopee Indonesia). *E-Proceeding of Management*, *6*(1), 12–19.

Kurniawan, S., Gata, W., Puspitawati, D. A., Nurmalasari, Tabrani, M., & Novel, K. (2019). Perbandingan Metode Klasifikasi Analisis Sentimen Tokoh Politik Pada Komentar Media Berita Online. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *3*(2), 176–183. https://doi.org/10.29207/resti.v3i2.935

Limbong, J. J. A., Sembiring, I., & Hartomo, K. D. (2022). Analisis Klasifikasi Sentimen Ulasan Pada E-Commerce Shopee Berbasis Word Cloud Dengan Metode Naive Bayes Dan K-Nearest Neighbor. *Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIIK)*, *9*(2), 347–356. https://doi.org/10.25126/jtiik.202294960

Muktafin, E. H., Kusrini, K., & Luthfi, E. T. (2020). Analisis Sentimen pada Ulasan Pembelian Produk di Marketplace Shopee Menggunakan Pendekatan Natural Language Processing. *Jurnal Eksplora Informatika*, *10*(1), 32–42. https://doi.org/10.30864/eksplora.v10i1.390

Negara, A. B. P., Muhardi, H., & Putri, I. M. (2020). Analisis Sentimen Maskapai Penerbangan Menggunakan Metode Naive Bayes dan Seleksi Fitur Information Gain. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *7*(3), 599. https://doi.org/10.25126/jtiik.2020711947

Nurdin, Hutomi, M., Qamal, M., & Bustami, B. (2020). Sistem Pengecekan Toko Online Asli atau Dropship pada Shopee Menggunakan Algoritma Breadth First Search. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *4*(6), 1117–1123. https://doi.org/10.29207/resti.v4i6.2514

Pratmanto, D., Rousyati, R., Wati, F. F., Widodo, A. E., Suleman, S., & Wijianto, R. (2020). App Review Sentiment Analysis Shopee Application in Google Play Store Using Naive Bayes Algorithm. *Journal of Physics: Conference Series*, *1641*(1). https://doi.org/10.1088/1742-6596/1641/1/012043

Rahat, A. M., Kahir, A., & Masum, A. K. M. (2019). Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset. *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, 266–270. https://doi.org/10.1109/SMART46866.2019.9117512

Rhohmawati, U., Slamet, I., & Pratiwi, H. (2019). Sentiment Analysis Using Maximum Entropy on Application Reviews (Study Case: Shopee on Google Play). *Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika (JITEKI)*, *5*(1), 44–49. https://doi.org/10.26555/jiteki.v5i1.13087

Rolliawati, D., Khalid, K., & Rozas, I. S. (2020). Teknologi Opinion Mining untuk Mendukung Strategic Planning. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *7*(2), 293. https://doi.org/10.25126/jtiik.2020721685

Saputra, F. T., Nurhadryani, Y., Wijaya, S. H., & Defina, D. (2021). Analisis Sentimen Bahasa Indonesia pada Twitter Menggunakan Struktur Tree Berbasis Leksikon. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *8*(1), 135. https://doi.org/10.25126/jtiik.0814133

Sihombing, L. O., Hannie, H., & Dermawan, B. A. (2021). Sentimen Analisis Customer Review Produk Shopee Indonesia Menggunakan Algortima Naïve Bayes Classifier. *Edumatic: Jurnal Pendidikan Informatika*, *5*(2), 233–242. https://doi.org/10.29408/edumatic.v5i2.4089

Siswanto, Wibawa, Y. P., Gata, W., Gata, G., & Kusumawardhani, N. (2018). Classification Analysis of MotoGP Comments on Media Social Twitter Using Algorithm Support Vector Machine and Naive Bayes. *2018 International Conference on Applied Information Technology and Innovation (ICAITI)*, 96–101. https://doi.org/10.1109/ICAITI.2018.8686751

Somantri, O., & Apriliani, D. (2018). Support Vector Machine Berbasis Feature Selection Untuk Sentiment

*name of corresponding author

Analysis Kepuasan Pelanggan Terhadap Pelayanan Warung dan Restoran Kuliner Kota Tegal. *Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIIK)*, *5*(5), 537–548. https://doi.org/10.25126/jtiik.201855867

Widyastuti, H., & Prastitya, T. A. (2020). Preferensi Konsumen Pengguna E-Commerce yang Memengaruhi Kesadaran akan Perlindungan Konsumen pada Generasi X. *Jurnal Sistem Informasi Bisnis*, *10*(1), 10–19. https://doi.org/10.21456/vol10iss1pp10-19

Wiratama, G. P., & Rusli, A. (2019). Sentiment Analysis of Application User Feedback in Bahasa Indonesia Using Multinomial Naive Bayes. *2019 5th International Conference on New Media Studies (CONMEDIA)*, 223–227. https://doi.org/10.1109/CONMEDIA46929.2019.8981850

*name of corresponding author