Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13559
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Comparison of Naïve Bayes and SVM in Sentiment Analysis of Product Reviews on Marketplaces

Nurul Zalza Bilal Jannah[1)*], Kusnawi[2)]
[1)2)3)]Universitas Amikom Yogyakarta, Yogyakarta, Indonesia
[1)]zalzabil@students.amikom.ac.id , [2)]khusnawi@amikom.ac.id

Submitted :Mar 4, 2024 | **Accepted** : Mar 21, 2024 | **Published** : Apr 1, 2024

Abstract: At this time more and more people are switching to shopping online in existing marketplaces such as Shopee. Marketplaces provide various advantages and disadvantages to customers such as lower costs and goods sent not according to orders. Product reviews from customers greatly affect the sales level of business people so that sentiment analysis is carried out. The importance of conducting sentiment analysis of product reviews in the marketplace is to add an overview of how the product is received by users. This research uses Naïve Bayes and SVM algorithms for sentiment analysis of beauty care product review datasets obtained from Shopee scraping results. This research implements k fold cross validation for data splitting process of 10 folds. The Naïve Bayes algorithm obtained the highest accuracy value of 85.53% on fold 2 and the lowest accuracy value of 77.16% on fold 3. While the SVM algorithm obtained the highest accuracy value of 88.58% on fold 2 and the lowest accuracy value of 82.99% on fold 7. With this it is stated that SVM can work better for sentiment analysis of beauty care product reviews on the Shopee marketplace because it gets a higher average accuracy value of 86.14% compared to the Naïve Bayes algorithm.

Keywords: Sentiment Analysis; Review Products; Naïve Bayes; SVM; Marketplaces

INTRODUCTION

In the last few years, online shopping has developed quite rapidly. More and more people are turning to shop online at existing marketplaces such as Shopee. Various kinds of products sold in the marketplace, one of which is beauty care products that are widely discussed among women and men. Beauty care products include lipstick, body lotion & body butter, facial wash, powder, eyebrow, etc. Customers will provide reviews on the products they buy in the form of positive and negative reviews where product reviews are one source of information about product quality and are very influential on consumers (Wang & Wang, 2020).

In this context, product reviews from customers greatly affect the sales level of business people. It is important to conduct sentiment analysis of product reviews in the marketplace to add an overview of how the product is received by users. Sentiment analysis is a technique used to extract and understand the opinions that exist in a review in the form of text (Wahyudi & Sibaroni, 2022).

This research uses Naïve Bayes algorithm and SVM for sentiment analysis of product reviews. The Naïve Bayes algorithm utilizes Bayes' theorem to calculate the probability of a particular sentiment based on features present in text reviews. Naïve Bayes can also be called idiot's Bayes, simple Bayes, and independence Bayes because it is easy to apply without requiring a complex iteration process (Sri Diantika et al., 2021). SVM is an appropriate algorithm for text classification with its ability to find the best hyperplane making this algorithm has a high level of generality and good accuracy (Nasution & Hayaty, 2019).

Based on research (Fikri et al., 2020) the comparison between the naive bayes and SVM methods on twitter sentiment analysis resulted in a Naive Bayes accuracy value of 73.65% and an SVM accuracy value of 70.20%. The research proves that Naive Bayes is able to work better than SVM. Therefore, this research will compare Naive Bayes and SVM by adding parameters to the modeling algorithm and using cross validation for split data.

This research uses Naïve Bayes and SVM algorithms for sentiment analysis of beauty care product review datasets obtained from Shopee scraping results. The dataset is cleaned by eliminating reviews that do not match the research objectives. The data is processed by labeling positive or negative sentiments using the vader lexicon. Furthermore, it goes through preprocessing stages (case folding, cleansing, word normalization, stopword removal,

*name of corresponding author

stemming, and tokenizing). Then divided using the K-Fold Cross Validation method followed by TF-IDF and data modeling by adding parameters to improve accuracy. Model evaluation using confusion matrix to produce classification report in the form of accuracy, precision, recall, and F1 score.

LITERATURE REVIEW

Literature review is very important to be done as a reference material in a study. Here are some references used in this research. Research by (Fide et al., 2021) produced an SVM kernel RBF algorithm with a ratio of training data and test data of 80: 20 and the best accuracy rate of 90.62% and kappa of 81.24% so that it is included in the almost perfect classification results. Based on the classification and association results, users give positive reviews because they like and are comfortable with TikTok which contains funny videos on fyp. While users who give negative reviews because they fail to register and their accounts are blocked, so users ask for improvements from TikTok. In research (Alfiah Zulqornain & Pandu Adikara, 2021) produces an accuracy value of 0.729947, a precision value of 0.746854, a recall value of 0.926118, and an f-measure value of 0.824511 by using the Naïve Bayes, Categorial Propotional Difference, and 5-Cross Validation methods with term variations.

In research (Maodah et al., 2023) obtained a combination of the third preprocessing technique (case folding, punctuation removal, word normalizer, and stemming), a combination of the second word2vec parameters (size 50, window 2, hs 0, and negative 10), and a combination of the fourth CNN parameters (kernel size 2, dropout 0. 2, and learning rate 0.01) has the best accuracy of 99.00%, precision of 98.96%, and recall of 98.9%. Research (Wahyudi & Sibaroni, 2022) explains that sentiment classification is carried out on each predetermined aspect, namely aspects of features, business, and content. The deep learning method RNN-LSTM and the addition of word embedding BERT produce the highest sentiment classification with an accuracy of 0.94 obtained from business aspects, an accuracy of 0.91 obtained from content aspects, and feature aspects with the lowest accuracy of 0.85. Research (Kosasih & Alberto, 2021) explains that the data collected is 1000 reviews divided into 700 training data and 300 testing data. The reviews are processed through case folding, tokenizing, stopwords, stemming, and weighting words with TF-IDF. Based on the classification results, the accuracy rate is 80.2223% and the F1 score is 0.691372.

Based on research (Sihombing et al., 2021) explains that by applying the Knowledge Discovery in Text (KDT) methodology and the Naïve Bayes Classifier algorithm to extract information from text data, it produces an accuracy value of 85% which can be used as an evaluation material for business people. Then the research (Kusnawi et al., 2023) explains that the research aims to classify Neobank user review data, including positive or negative sentiment. The method used in this research is an experimental method with the SVM algorithm managed to get the highest accuracy value of 82.33%, owned by a scenario of 90% training data and 10% testing data. This research has a precision value of 82% and recall 81%.

METHOD

In conducting this research, the stages of the research were divided into 10 stages which can be seen in this figure.

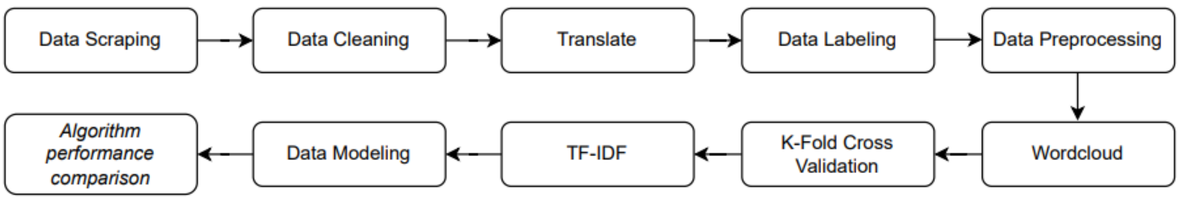

Fig 1. Research Flow

Data Scraping

The data collection in this research employed the scraping method from the Shopee marketplace, involving the retrieval of semi-structured documents from the internet (Mufidah, 2021). Subsequently, these web pages were analyzed to extract specific data that could be utilized in other contexts. The data obtained from the scraping process was then stored in a database and could be exported in CSV format.

Data Cleaning

In this research, the data cleaning process was carried out through several stages, including eliminating empty data, removing data without alphabets, and deleting duplicate data. The study, which focuses on product reviews, involved the elimination of certain inappropriate data, such as the removal of reviews related to shipping, packaging, and store service.

*name of corresponding author

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13559
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Translate

The dataset was then translated from Bahasa Indonesia to English for labeling with the VADER lexicon, which is specifically designed for the English language (Sumitro et al., 2021). In this study, Google Translator in the Python library will be used to translate the dataset.

Data Labelling

This research employs the lexicon method for data labeling using the VADER dictionary designed specifically to detect sentiment in English text, containing 7,500 tokens. After going through the translation stage, VADER will function by invoking lexicon data from the NLTK server to calculate the polarity class of sentiment (Abimanyu et al., 2022).

Data Preprocessing

In the case folding stage, the entire text is transformed into lowercase. In the cleansing stage, non-ASCII characters, emoticons, numbers, punctuation, single characters, and white spaces are removed. Next is the word normalization stage, which is used to correct non-standard words and language. In this research, a word normalizer utilizing prosa.ai as an NLP API source for processing Indonesian text is employed. The stopword removal stage involves eliminating unimportant words that are considered to provide no meaningful information, such as 'di' (in), 'yang' (which), 'dan' (and), 'atau' (or), and others (Rabbani et al., 2023). The subsequent step is stemming, where affixes are removed from words in the sentence, reducing them to their basic forms. The stemming process in this study uses the Sastrawi library (Maodah et al., 2023). Following that is tokenization, breaking the text into individual tokens. A unique word list is then created from the text, serving as a numerical representation. These tokens are converted into numerical representations according to the vocabulary.

Wordcloud

Wordcloud is a data visualization technique employed in this research with the aim of displaying frequently occurring words in reviews. The visualization of words presented in this stage consists of two sentiments: positive and negative sentiments.

K-Fold Cross Validation

The data splitting stage in this research utilizes the k-fold cross-validation method, aiming to check for overfitting in a model (Nasution & Hayaty, 2019). This study employs 10-fold cross-validation to divide the dataset into 10 different parts, where each iteration involves 9 parts as the training set and 1 part as the validation set.

TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) is a technique used for variable reduction in the feature extraction stage for the dataset being utilized (Sidik et al., 2022). This word weighting aims to assign weights to word features based on their frequency of occurrence. TF represents the frequency of a word in a document, while IDF is a weighting factor used to determine whether the sought term matches the desired keywords.

Data Modeling

The first step in data modeling is training the algorithms using libraries. The library used is MultinomialNB for the Naïve Bayes algorithm and SVC for the SVM algorithm, with added parameters. The parameter for Naive Bayes is alpha=0.1 to control the smoothing effect applied, and fit_prior set to False to make the model consider all classes equally. Meanwhile, the parameter for SVM is kernel='linear', which is suitable for handling problems with linear decision boundaries. The next step is to evaluate the trained model using cross-validation, iterating over 10 folds to predict on the testing data. These predictions are then compared with the actual labels in the testing data by calculating accuracy using the confusion matrix method. The conclusion of this data modeling is to display the classification results of the algorithms in 10-fold cross-validation using a classification report.

Naïve Bayes

Naïve Bayes is the most commonly used methodology for sentiment analysis. This algorithm uses predictions based on Bayes' theorem. This method belongs to the supervised learning class for classification (Rahmadani et al., 2022). This algorithm uses conditional probability to calculate the probability of an event being assigned to a particular class (Kosasih & Alberto, 2021). For its classification domain, the computed probability is P(H|X), which is the likelihood that the hypothesis is true for the observed sample data X, and it can be applied to the following formula.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \qquad (1)$$

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13559
e-ISSN : 2541-2019
p-ISSN : 2541-044X

SVM

SVM operates by searching for an optimal hyperplane that provides the maximum distance between the two classes. Margin is measured as the distance between the closest data points to the hyperplane and these points are referred to as support vectors (Tineges et al., 2020). SVM utilizes a kernel function that transforms data into a higher-dimensional space, intending to enhance the data structure for easier separation (Fide et al., 2021). The following is the general formula for linear SVM:

$$f(x) = sign\ (w.x + b) \qquad (2)$$

Algorithm Performance Comparison

The performance comparison of these algorithms is carried out by comparing the modeling results of Naïve Bayes and SVM algorithms in each fold of k-fold cross-validation. This involves calculating the average performance results of the models in each fold.

RESULT

The dataset to be used consists of 7,163 records. Afterward, data cleaning is performed to eliminate empty data, duplicates, and reviews that do not align with the research objectives, resulting in 3,940 records. The data is then processed through preprocessing steps (case folding, cleansing, word normalization, stopword removal, stemming, and tokenizing). After labeling the data with lexicons, visualization is conducted to display the occurrence of words with positive and negative sentiments.

Fig 2. Wordcloud Positive Data

Fig 3. Wordcloud Negative Data

The dataset is then divided through the k-fold cross-validation process with 10 folds, and the results are obtained as follows.

Table 1 10-Fold Cross Validation Results

Fold	Training Sample Negatif	Training Sample Positif	Testing Sample Negatif	Testing Sample Positif
1	846	2700	93	301

*name of corresponding author

2	845	2701	94	300
3	845	2701	94	300
4	845	2701	94	300
5	845	2701	94	300
6	845	2701	94	300
7	845	2701	94	300
8	845	2701	94	300
9	845	2701	94	300
10	845	2701	94	300

Feature extraction using TF-IDF and its results for algorithm modeling. Parameters are added to the algorithm modeling, and the evaluation employs a confusion matrix. Subsequently, accuracy, precision, recall, and F1 score are calculated for display in a classification report, as shown in the following table.

Table 2 NB Confusion Matrix Results

	Accuracy	*Precision*	*Recall*	*F1-Score*
Fold 1	0,80	0,91	0,81	0,86
Fold 2	0,86	0,94	0,86	0,90
Fold 3	0,77	0,92	0,76	0,84
Fold 4	0,82	0,90	0,86	0,88
Fold 5	0,83	0,92	0,85	0,88
Fold 6	0,81	0,91	0,82	0,87
Fold 7	0,80	0,90	0,82	0,86
Fold 8	0,81	0,91	0,83	0,87
Fold 9	0,82	0,95	0,81	0,87
Fold 10	0,83	0,94	0,83	0,88

Table 3 SVM Confusion Matrix Results

	Accuracy	*Precision*	*Recall*	*F1-Score*
Fold 1	0,88	0,90	0,94	0,92
Fold 2	0,89	0,91	0,95	0,93
Fold 3	0,86	0,90	0,91	0,91
Fold 4	0,87	0,88	0,96	0,92
Fold 5	0,84	0,88	0,92	0,90
Fold 6	0,86	0,90	0,92	0,91
Fold 7	0,83	0,85	0,94	0,89
Fold 8	0,86	0,88	0,94	0,91
Fold 9	0,88	0,90	0,94	0,92
Fold 10	0,87	0,89	0,93	0,91

The accuracy results for each fold will be averaged to compare which algorithm performs better, as illustrated in the following figure.

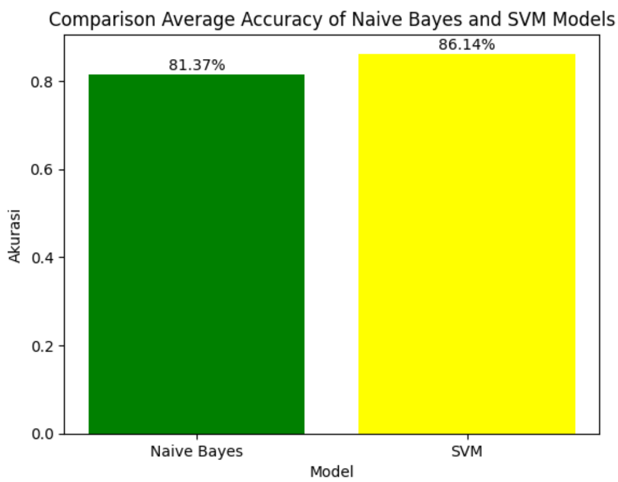

Fig 4. Comparison of NB and SVM Accuracy at Each Fold

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13559
e-ISSN : 2541-2019
p-ISSN : 2541-044X

DISCUSSIONS

From the results of the conducted research, at this stage, an evaluation is performed to determine the differences between modeling algorithms with added parameters and without added parameters. After adding parameters to each algorithm, higher accuracy values were obtained for each fold. The following is a comparison of the average accuracy values between modeling the Naïve Bayes and SVM algorithms with parameters and without parameters.

Table 4 Comparison Algorithms with Parameters and without Parameters

	Naïve Bayes	*SVM*
With Parameter	81,37%	86,14%
Without Parameter	77,17%	85,53%

The table explains that modeling the algorithm by adding parameters results in better accuracy compared to modeling the algorithm without parameters. This study is somewhat similar to the previous research (Fikri et al., 2020), but the earlier study compared Naïve Bayes and SVM without parameters. In that research, the accuracy for Naïve Bayes was found to be 73.65%, and the accuracy for SVM was 70.20%.

CONCLUSION

The Naïve Bayes algorithm obtained the highest accuracy value of 85.53% on fold 2 and the lowest accuracy value of 77.16% on fold 3. While the SVM algorithm obtained the highest accuracy value of 88.58% on fold 2 and the lowest accuracy value of 82.99% on fold 7. The SVM algorithm has better performance for sentiment analysis of beauty care product reviews on the Shopee marketplace because it obtained a higher average accuracy value of 86.14% compared to the Naïve Bayes algorithm which only obtained an average accuracy value of 81.37%. The results of sentiment analysis with a high enough accuracy rate can help customers to add an overview and insight into how the product is received by users, as well as help sellers to improve the products they sell, get positive and negative feedback, and also be able to overcome problems that will arise when getting bad reviews in the future.

REFERENCES

Abimanyu, D., Budianita, E., Cynthia, E. P., Yanto, F., & Yusra, Y. (2022). Analisis Sentimen Akun Twitter Apex Legends Menggunakan VADER. *Jurnal Nasional Komputasi Dan Teknologi Informasi (JNKTI)*, *5*(3), 423–431. https://doi.org/10.32672/jnkti.v5i3.4382

Alfiah Zulqornain, J., & Pandu Adikara, P. (2021). *Analisis Sentimen Tanggapan Masyarakat Aplikasi Tiktok Menggunakan Metode Naïve Bayes dan Categorial Propotional Difference (CPD)*. *5*(7), 2886–2890. http://j-ptiik.ub.ac.id

Fide, S., Suparti, & Sudarno. (2021). *ANALISIS SENTIMEN ULASAN APLIKASI TIKTOK DI GOOGLE PLAY MENGGUNAKAN METODE SUPPORT VECTOR MACHINE (SVM) DAN ASOSIASI*. *10*, 346–358.

Fikri, M. I., Sabrila, T. S., & Azhar, Y. (2020). Perbandingan Metode Naïve Bayes dan Support Vector Machine pada Analisis Sentimen Twitter. *Smatika Jurnal*, *10*(02), 71–76. https://doi.org/10.32664/smatika.v10i02.455

Kosasih, R., & Alberto, A. (2021). Sentiment analysis of game product on shopee using the TF-IDF method and naive bayes classifier. *ILKOM Jurnal Ilmiah*, *13*(2), 101–109. https://doi.org/10.33096/ilkom.v13i2.721.101-109

Kusnawi, Rahardi, M., & Pandiangan, V. D. (2023). Sentiment Analysis of Neobank Digital Banking Using Support Vector Machine Algorithm in Indonesia. *International Journal on Informatics Visualization*, *7*(2), 377–383. https://doi.org/10.30630/joiv.7.2.1652

Maodah, F., Utami, E., & ... (2023). Optimizing Sentiment Analysis of Product Reviews on Marketplace Using a Combination of Preprocessing Techniques, Word2Vec …. *Jurnal Teknik Informatika …*, *4*(1), 1–7. http://jutif.if.unsoed.ac.id/index.php/jurnal/article/view/815%0Ahttp://jutif.if.unsoed.ac.id/index.php/jurnal/article/download/815/268

Mufidah, U. (2021). Perancangan Aplikasi Perbandingan Harga Produk (Historical Data) Menggunakan Teknik Web Scraping. *Pusdansi.Org*, *1*(1), 1–14.

Nasution, M. R. A., & Hayaty, M. (2019). Perbandingan Akurasi dan Waktu Proses Algoritma K-NN dan SVM dalam Analisis Sentimen Twitter. *Jurnal Informatika*, *6*(2), 226–235. https://doi.org/10.31311/ji.v6i2.5129

Rahmadani, P. S., Tampubolon, F. C., Jannah, A. N., Hutabarat, N. L. H., & Simarmata, A. M. (2022). Tiktok Social Media Sentiment Analysis Using the Nave Bayes Classifier Algorithm. *SinkrOn*, *7*(3), 995–999. https://doi.org/10.33395/sinkron.v7i3.11579

Sidik, F., Suhada, I., Anwar, A. H., & Hasan, F. N. (2022). Analisis Sentimen Terhadap Pembelajaran Daring Dengan Algoritma Naive Bayes Classifier. *Jurnal Linguistik Komputasional (JLK)*, *5*(1), 34. https://doi.org/10.26418/jlk.v5i1.79

Sihombing, L. O., Hannie, H., & Dermawan, B. A. (2021). Sentimen Analisis Customer Review Produk Shopee

*name of corresponding author

Indonesia Menggunakan Algortima Naïve Bayes Classifier. *Edumatic: Jurnal Pendidikan Informatika*, *5*(2), 233–242. https://doi.org/10.29408/edumatic.v5i2.4089

Sri Diantika, Windu Gata, & Hiya Nalatissifa. (2021). Komparasi Algoritma SVM Dan Naive Bayes Untuk Klasifikasi Kestabilan Jaringan Listrik. *Elkom : Jurnal Elektronika Dan Komputer*, *14*(1), 10–15. https://doi.org/10.51903/elkom.v14i1.319

Sumitro, P. A., Rasiban, Mulyana, D. I., & Saputro, W. (2021). Analisis Sentimen Terhadap Vaksin Covid-19 di Indonesia pada Twitter Menggunakan Metode Lexicon Based. *J-ICOM - Jurnal Informatika Dan Teknologi Komputer*, *2*(2), 50–56. https://doi.org/10.33059/j-icom.v2i2.4009

Tineges, R., Triayudi, A., & Sholihati, I. D. (2020). Analisis Sentimen Terhadap Layanan Indihome Berdasarkan Twitter Dengan Metode Klasifikasi Support Vector Machine (SVM). *Jurnal Media Informatika Budidarma*, *4*(3), 650. https://doi.org/10.30865/mib.v4i3.2181

Wahyudi, D., & Sibaroni, Y. (2022). Deep Learning for Multi-Aspect Sentiment Analysis of TikTok App using the RNN-LSTM Method. *Building of Informatics, Technology and Science (BITS)*, *4*(1), 169–177. https://doi.org/10.47065/bits.v4i1.1665

Wang, H., & Wang, Y. (2020). A Review of Online Product Reviews. *Journal of Service Science and Management*, *13*(01), 88–96. https://doi.org/10.4236/jssm.2020.131006

*name of corresponding author