BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 3, No 2, Juni 2022, Hal 115 - 120
ISSN 2722-0524 (media online)
DOI 10.47065/ bit.v3i1.282
https://journal.fkpt.org/index.php/BIT

A Implementation of Text Mining In Sentiment Analysis of Shopee Indonesia Using SVM

Kusdarnowo Hantoro, Dwipa Handayani, Siti Setiawati
Computer Science, Department of Informatics, Bhayangkara Jaya University, Indonesia
Email: [1]kusdarnowo@dsn.ubharajaya.ac.id, [2]dwipa.handayani@dsn.ubharajaya.ac.id, [3]sitisetiawati@dsn.ubharajaya.ac.id
Email Penulis Korespondensi: *kusdarnowo@dsn.ubharajaya.ac.id

Abstract− Many online shopping users convey an assessment of a product through status comments. Comments indicate a non-standard form of expression. Nowadays, user comments are increasing rapidly, which makes data management difficult. Today's society has so many choices; most of them are used for preference as a final recommendation. The top item shows the preferences of the available items, recommended based on future predictability. The predictive rating of some items is used by the user to recommend the item to other users. Sentiment analysis can recommend items of choice for online shopping in various fields of application, especially in e-commerce. Sentiment analysis is used to identify opinions, ideas, or thoughts from online media. Shopee is one of the largest online business platforms. Classification Sentiment analysis towards Shopee based on the Support Vector Machine (SVM) was used in this study on the Shopee application using 990 training data and 110 test data. From the test data, 28 data entered the negative class and the remaining 82 data entered the positive class and resulted in an accuracy rate of 80.90%, meaning that from 110 assessments there were 89 assessments classified exactly in the sentiment class.

Keywords: Online Business; Opinion; Assessment; Classification; Recommendation.

Abstrak− Banyak pengguna belanja online menyampaikan penilaian suatu produk melalui komentar status. Komentar menunjukkan bentuk ekspresi yang tidak standar. Saat ini, komentar pengguna meningkat pesat, yang membuat pengelolaan data menjadi sulit. Masyarakat saat ini memiliki begitu banyak pilihan; kebanyakan dari mereka digunakan untuk preferensi sebagai rekomendasi akhir. Item teratas menunjukkan preferensi item yang tersedia, yang direkomendasikan berdasarkan prediktabilitas di masa mendatang. Peringkat prediktif dari beberapa item digunakan oleh pengguna untuk merekomendasikan item tersebut kepada pengguna lain. Analisis sentimen dapat merekomendasikan item pilihan untuk belanja online di berbagai bidang aplikasi, terutama di e-commerce. Analisis sentimen digunakan untuk mengidentifikasi opini, ide, atau pemikiran dari media online. Shopee adalah salah satu platform bisnis online terbesar. Analisis klasifikasi Sentimen terhadap Shopee berdasarkan Support Vector Machine (SVM) digunakan dalam penelitian ini pada aplikasi Shopee dengan menggunakan 990 data latih dan 110 data uji. Dari data pengujian, 28 data masuk ke kelas negatif dan sisanya 82 data masuk ke kelas positif dan menghasilkan tingkat akurasi 80,90%, artinya dari 110 penilaian terdapat 89 penilaian yang tergolong tepat pada kelas sentimen

Kata Kunci: Bisnis Daring, Opini, Penilaian, Klasifikasi, Rekomendasi.

1. INTRODUCTION

The rapid growth of online stores has driven intense competition among e-commerce sales business. E-commerce causes changes in consumer behavior in line with changing market conditions and increasingly competitive competition. In Indonesia, online shopping is the choice of many parties in buying goods [1]. The continued growth of e-commerce in Indonesia has enabled Shopee to activate this industry. According to "Map E-Commerce Indonesia" data released by iPrice, Shopee is the market with the largest number of visitors in Indonesia in the first quarter of 2022, with a monthly number of visitors reaching 8 million. If Shopee can provide the right service according to consumer expectations, then Shopee will have a good perception in the eyes of consumers. In order to provide appropriate and appropriate services, companies must understand consumer expectations and provide satisfactory service [2]. If consumers are satisfied with the services provided, consumers tend to compare them with the services of other companies.

Sentiment analysis is a study to assess people's attitudes, emotions, and opinions based on text expressions, involving a combination of text mining and natural language processing (NLP) [3]. Sentiment analysis is widely applied in various fields such as politics (to predict election results from political forums), business (to analyze online sentiment on social media for stock market predictions) and marketing (to forecast sales of certain products)[4][5][6]. In conducting the analysis, we assume that the document contains an opinion. However, in most cases, only the objective information and facts stated in these documents (one example being news documents) are used. Sometimes, documents that include factual sentences are considered to contain a sentiment (opinion) section. Therefore, identifying the type and nature of the sentence is the most basic part of sentiment analysis[7][4][8]. Analysis is given to the sentences used and then extracted based on subjective or objective analysis. Classification of subjectivity is the main task in sentiment analysis, which involves classifying sentences as objective or subjective.

In conducting sentiment analysis on sentences, you can use dictionary-based, machine learning or hybrid methods. Dictionary-based sentiment analysis uses dictionaries that are labeled with values. And machine learning approaches generally use text classification algorithms to identify patterns in text. The approach with the hybrid method combines the two methods[9].

The machine learning approach uses several techniques in an attempt to extract salient features more accurately to provide information about sentiment polarity. One of the classification algorithms that is often used and gets a lot of attention from researchers is Support Vector Machines. Sentiment classification includes the function of checking the sentiment polarity of the sentences that have been filtered. These sentences are categorized into neutral, negative or

BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 3, No 2, Juni 2022, Hal 115 - 120
ISSN 2722-0524 (media online)
DOI 10.47065/ bit.v3i1.282
https://journal.fkpt.org/index.php/BIT

positive sentiment depending on the case. Sentiment analysis in sentences has a bigger challenge, especially in local languages (non-English). There are several reasons, first, the supply of English text data is richer than other text data. Second, the process of compiling English text data is clearer, such as library stemming, lemmatization, and stop-word removal.

2. METHODOLOGY

2.1 Text Pre-processing

Application review data on the website (https://play.google.com/store/apps/details?id=com.shopee.id) is unstructured text data, text data will be processed using text mining methods to obtain information that can be useful for various parties. This process is the stage carried out to transform data from unstructured form into structured data so that it is easy to analyze further. The steps taken include case folding, remove punctuation, filtering and stemming

2.2 Case Folding

The process of changing letters into one form in a text document, for example changing from capital to lowercase or vice versa. So this process is a uniform form of data that will be processed in one form. The process of uniform data form is to change the text data into lower case letters. The letter "A" in the word "Aplikasi", will be converted into lower case letters to match the shape of the text to be analyzed so that the result becomes "aplikasi".

2.3 Remove Punctuation

The process by which the system removes punctuation marks or symbols in the dataset. These punctuation marks or symbols are removed because they have no effect on the results of sentiment analysis.
Example:

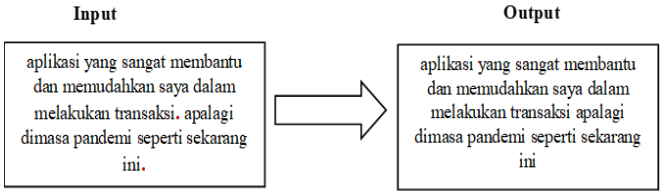

Figure 1. Remove Punctuation

2.4 Stop words removal

The filtering or filtering stage is the stage of selecting words in the document to reduce the dimensions of sentences in the corpus called stop words. Stopwords is a step to eliminate words that are not influential / not informative but often appear in the document. Some words that will be omitted are conjunctions, pronouns, prepositions, unwanted words.

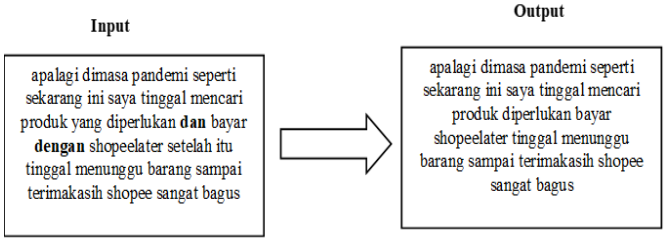

Figure 2. Stop words removal

2.5 Stemming

Stemming means removing prefixes and suffixes to produce a root word. This process is also known as conflation. The stemming process is widely used in Information retrieval (information seeking) to improve the quality of information that can be obtained.

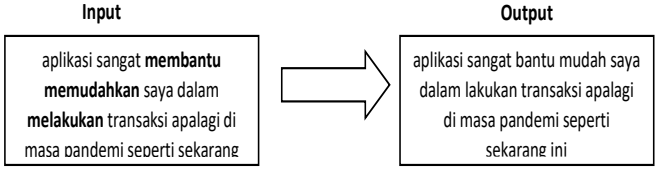

Figure 3. Stemming

BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 3, No 2, Juni 2022, Hal 115 - 120
ISSN 2722-0524 (media online)
DOI 10.47065/ bit.v3i1.282
https://journal.fkpt.org/index.php/BIT

2.6 Simulation

Based on user review text testing: "aplikasi sangat bantu mudah saya dalam lakukan transaksi apalagi dimasa pandemi seperti sekarang ini saya tinggal cari produk diperlukan bayar shopeelater setelah itu tinggal tunggu barang sampai terimakasih shopee sangat bagus", a simulation is carried out. The results obtained 4 positive words and 0 detected negative words, namely "bantu", "mudah", "terima kasih", "bagus", as positive words. The formula for calculating the sentiment score used in the labeling process is as follows:

Table 1. Simulation

Review	Positive	Negative
aplikasi sangat **bantu mudah** saya dalam lakukan transaksi apalagi dimasa pandemi seperti sekarang ini saya tinggal cari produk diperlukan bayar shopeelater setelah itu tinggal tunggu barang sampai **terimakasih** shopee sangat **bagus**	bantu mudah terimakasih bagus	-
Number	4	0
Calculation	Score = 4-0 Score = 4	

Thus, the following calculation is obtained:
Score = (Number of positive words) - (Number of negative words)
Score = 4-0=4
The final score obtained from the calculation simulation is > 0, so the results of the review classification are positive.

2.7 Sentiment Class Labelling

Support Vector Machines (SVM) is a machine learning algorithm capable of handling multiple variables and classes. SVM is a machine learning method that uses generalization theory train linear machine learning efficiently on induced kernel features. SVM tries to reduce the possibility of misclassification of test data that is not visible to the model and retrieved in random from a known but unknown probability distribution[10][11]

After going through the preprocessing process, sentiment analysis will then be carried out for data labeling. The data labeling process is carried out automatically by the Lexicon dictionary by calculating a sentiment score[11][12][13]. Word weighting is done by calculating the frequency of occurrence of words in a text document. The more often a word appears in a text document, the greater the weight of the word and the word is considered a word that strongly represents the text document[14].

In general, sentiment analysis is used to classify (label) text documents into three classes of sentiment, namely positive, neutral and negative sentiment[15]. The way to determine the sentiment class is to calculate the score for the number of positive words minus the score for the number of negative words in each review sentence[12]. Sentences that have a score of > 0 will be classified into positive class, sentences that have a score of = 0 will be classified into neutral class, while sentences that have a score of < 0 are classified into negative class. However, in this study, two sentiment class labels were used, namely positive sentiment and negative sentiment. The result of text classification is a classification model. The model will be tested against training data and test data and then evaluated using several parameters. The evaluation metrics consist of: accuracy, precision, recall and scores are calculated using a confusion matrix (Table 1)

Table 2. Confusion Matrix

		Predict	
		Positive	Negative
Actual	Positive	TP (True Positive)	FP (False Positive)
	Negative	FN (False Negative)	TN (True Negative)

The percentage of events that are correctly classified is called accuracy, which is indicated in the equation. The most intuitive performance metric is accuracy, which is the ratio of correctly estimated observations to all observations. As shown in equation 1 :

$$Accuracy = \frac{TP + FN}{TP + FP + TN + FN} x100\%$$

Precision is the ratio of the number of related text documents identified in all text documents selected by the system. As shown in equation 2 :

BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 3, No 2, Juni 2022, Hal 115 - 120
ISSN 2722-0524 (media online)
DOI 10.47065/ bit.v3i1.282
https://journal.fkpt.org/index.php/BIT

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} x100\%$$

Recall is the ratio of the number of related text documents identified among all related text documents in the collection. As shown in equation 3:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} x100\%$$

3. RESULT

3.1 Descriptive Analysis

The descriptive analysis in this study is based on user comment data from the Google Play website to see an overview of Shopee application information from several aspects including the number of user ratings on the application, and the comparison of the number of comments by users of this application is divided into two categories, namely positive comments and negative comments. . Below is a rating that describes how users rate Shopee e-commerce on the Google Play website.

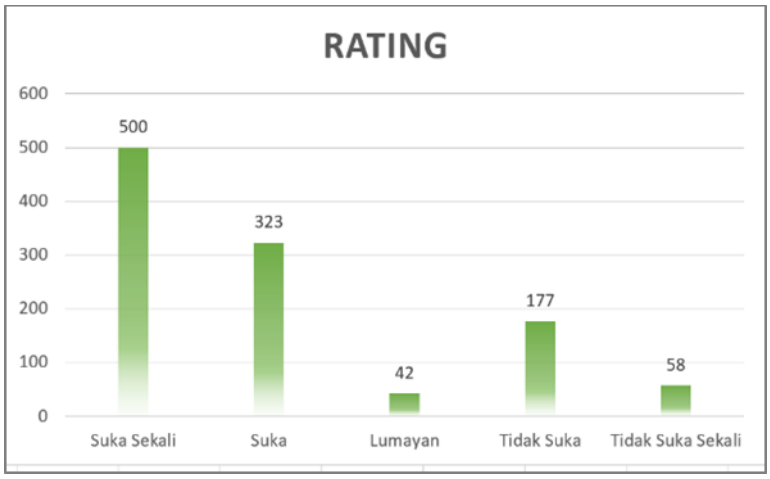

Figure 4. Shopee rating

Ratings on the Google Play site have a value of 1-5 with the lowest category being "Tidak Suka Sekali" which was given a score of "1","Tidak suka" with a score of "2","Lumayan" with a score of "3","Suka " with a score of "4", and "Suka Sekali" with a score of "5". From the picture above, it can be seen that the majority of Shopee users have good reviews of e-commerce. However, many Shopee users are out of sync between the content of the comments and the rating they provide.

3.2 Sentiment Class Labeling

The results of sentiment class labeling obtained the following data:

Table 1. Sentiment Class Labeling

Sentiment	Review
Positive	823
Negative	277

From the data training process, a classification model of each machine learning will be obtained. The model will then be tested to determine the level of model accuracy or the extent to which the model can classify the test data, this process is known as machine learning. Positive training data and negative training data are used by the SVM algorithm in studying data patterns based on the characteristics of the data in each class. The results of each machine learning experiment using the Support Vector Machine method are as follows:

Based on the table above, the results of sentiment class labeling show that the number of positive reviews has a higher frequency than the number of negative reviews and neutral reviews. The number of positive reviews is 823 reviews, neutral reviews are 82 reviews and negative reviews are 277 reviews.

The classification that will be used in this study is data with positive and negative sentiments. A review is classified as positive sentiment if it contains a positive statement such as a compliment, an expression of gratitude, or

BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 3, No 2, Juni 2022, Hal 115 - 120
ISSN 2722-0524 (media online)
DOI 10.47065/ bit.v3i1.282
https://journal.fkpt.org/index.php/BIT

a positive testimonial about Shopee e-commerce. A review is classified as negative sentiment if it contains negative statements such as dissatisfaction, insults, reports of service failures, and so on.

3.3 Training Data and Test Data

The training data is used by the classification algorithm to form a classifier model, this model is a knowledge representation that will be used to predict new data classes that have never existed, the larger the training data used, the better the machine will understand data patterns. Test data is used to measure the extent to which the classifier has successfully classified correctly. The data used for training data and test data is data that already has a class label, with the amount of training data and test data having a comparison These comparisons can be seen in the table below.

Table 2. Comparison of training data and test data

	Positive (823)		Negative (277)	
Percentage	Train	Test	Train	Test
90 : 10	741	82	249	28
80 : 20	658	165	222	55
70 : 30	576	247	194	83
60 : 40	494	329	166	111

Table 3. Scenario

Machine Learning	Data Composition	Accuracy
1	90:10	**80,90%**
2	80:20	78,77%
3	70:30	75,80%
4	60:40	76,68%

Based on the table above, of the 4 machine learning experiments conducted using the SVM method, machine learning 1 produced the highest level of accuracy, which was 80.90%. The results of the calculation of the level of accuracy obtained from the number of test data that are classified correctly compared to the total of all the data tested. To test the performance of the machine in classifying, cross validation was carried out using 4-fold cross validation with an average accuracy of 80.90%.

The confusion matrix is used to facilitate the accuracy calculation process by knowing the amount of test data that is classified correctly and the amount of test data that is misclassified. The comparison of the confusion matrix of the two methods obtained in machine learning can be seen in the following table:

Table 4. SVM Predictive Result

		Actual	
		Positive	Negative
Predict	Positive	68	14
	Negative	7	21
	Accuracy	80,90%	

The SVM method obtained predictive results that in the positive class, from 82 positive reviews tested, there were 68 reviews that had been classified correctly and there were prediction errors of 14 reviews that were included in negative reviews. While the negative reviews tested, out of a total of 28 reviews there were 21 reviews that were correctly classified as negative reviews and there were prediction errors as many as 7 reviews were included in positive reviews. Then from the value of the confusion matrix, an accuracy rate of 80.90% is obtained, meaning that from the 110 review data tested, there are 89 reviews that have the correct classification.

BULLETIN OF INFORMATION TECHNOLOGY (BIT)
Vol 3, No 2, Juni 2022, Hal 115 - 120
ISSN 2722-0524 (media online)
DOI 10.47065/ bit.v3i1.282
https://journal.fkpt.org/index.php/BIT

4. CONCLUSION

Based on the results of the analysis and discussion of the Shopee application review data, several conclusions were obtained as follows:

App rating in January 2022 – March 2022, 50% of users really like it and 5% really don't like the Shopee app out of a total of 1010 reviews.

Sentiment classification results using the SVM method on the Shopee application using 990 training data and 110 test data. From the test data, 28 data entered the negative class and the remaining 82 data entered the positive class and resulted in an accuracy rate of 80.90%, meaning that from 110 reviews there were 89 reviews classified exactly in the sentiment class..

ACKNOWLEDGEMENT

We thank to Salsa for her help and contribution by preparing supplying some specific data related to the implementation and testing.

REFERENCES

[1] C. S. M A Rahman, Aminul Islam, Esha B H, Sultana N, "Q2 - Consumer buying behavior towards online shopping_.pdf." pp. 1–22, 2018.
[2] M. Anisur Rahman, M. Aminul Islam, B. Humyra Esha, N. Sultana, and S. Chakravorty, "Cogent Business & Management Consumer buying behavior towards online shopping: An empirical study on Dhaka city, Bangladesh," Cogent Bus. Manag., vol. 5, p. 3, 2018.
[3] M. Khader, A. Awajan, and G. Al-Naymat, "The Effects of Natural Language Processing on Big Data Analysis: Sentiment Analysis Case Study," ACIT 2018 - 19th Int. Arab Conf. Inf. Technol., pp. 1–7, 2019.
[4] M. F. Daohai Zhang, Xue Liu, Juan Li, "Q0 - Sentiment Analysisv Enhanced Reader.pdf," in IOP Conf. Series: Journal of Physics: Conf. Series, 2018.
[5] A. F. Salam, H. Dai, and L. Wang, "Online Users' Identity Theft and Coping Strategies, Attribution and Sense of Urgency: A Non-Linear Quadratic Effect Assessment," Inf. Syst. Frointiers, 2021.
[6] M. Lu and F. Li, "Survey on lie group machine learning," Big Data Min. Anal., vol. 3, no. 4, pp. 235–258, 2020.
[7] D. Zhang et al., "Knowledge-oriented Hierarchical Neural Network for Sentiment Classification," in IOP Conference Series : Material Science and Engineering, 2019.
[8] A. V. Sakhare and P. S. Kasbe, "A review on road accident data analysis using data mining techniques," Proc. 2017 Int. Conf. Innov. Information, Embed. Commun. Syst. ICIIECS 2017, vol. 2018-Janua, pp. 1–5, 2018.
[9] I. M. Karo Karo, M. F. Md Fudzee, S. Kasim, and A. A. Ramli, "Sentiment Analysis in Karonese Tweet using Machine Learning," Indones. J. Electr. Eng. Informatics, vol. 10, no. 1, pp. 219–231, 2022.
[10] C. M. Liyew and H. A. Melese, "Machine learning techniques to predict daily rainfall amount."
[11] M. M. Rahman et al., "Sentiment topic mining based on comment tags."
[12] S. Jarp et al., "Supervised Ensemble Machine Learning Aided Performance Evaluation of Sentiment Classification," IOP Conf. Ser. J. Phys. Conf. Ser., vol. 1060, p. 12036, 2018.
[13] Y. Qi, H. Li, N. Liu, A. A. Arman, A. B. Kawi, and R. Hurriyati, "Sentiment analysis enhancement with target variable in Kumar's Algorithm," in IOP Conference Series : Material Science and Engineering, 2019, p. 128.
[14] D. S. Maylawati, Y. J. Kumar, and F. B. Kasmin, "Deep Learning-Now and Next in Text Mining and Natural Language Processing."
[15] S. Gao, "The Application of Information Classification in Agricultural Production Based on Internet of Things and Deep Learning," IEEE Access, vol. 10, pp. 22622–22630, 2022.