International Journal of Basic and Applied Science 11 (1) (2022) 37-46

Published by: IOCSCIENCE

International Journal of Basic and Applied Science

Journal homepage: www.ijobas.pelnus.ac.id

Performance Analysis Of Support Vector Machine In Identifying Comments And Ratings On E-Commerce

Mutiara S. Simanjuntak[1], Nurafni Damanik[2], Allwine[3]

[1]Department of Electrical and Computer Engineering, National Kaohsiung University Of Science and Tecnology, Jiangong Campus. No 415 Jiangong Rd., Sanmin Dist., Kaohsiung City 807, Taiwan, ROC

[23]Department of Electrical and Computer Engineering, National Taipei University of Technology, Zhongxiao E Rd, No. 1, Section 3, Da'an District, Taipei City, 106, Taiwan

Article Info

Article history:

Received Apr 28, 2022
Revised May 20, 2022
Accepted June 30, 2022

Keywords:

Confusion Matrix,
E-Commercem,
Google Colab Phyton,
Sentiment Analysis,
Support Vector Machine.

ABSTRACT

Consumers who have shopped at E-Commerce will provide reviews/comments on products that have been purchased. Customer confidence in the rating is hampered due to inconsistency of answers such as reviews that have negative text with a positive rating value. For this reason, a technique is needed to adjust the rating with comments or reviews of purchased goods to make it easier for consumers when shopping to see the rating directly without reading the reviews/comments of previous buyers. purpose of this study is to classify comments and ratings and then obtain the results of the accuracy of the classification system so that the above problems can be answered.This study uses Support Vector Machine classification technique because this algorithm is better in classification's terms. Data used are 1044 comment data and 1044 rating. Data are grouped into Good, Neutral, Less good categories using Python by Google Colab and divided into training and test data. To test capability of system, data that has been classified then analyzed using Confusion matrix. Results showed that SVM Algorithm was able to classify with an accuracy rate of 71.14%, 88% precision, and 79% recall.SVM algorithm is able to formulate training data with an accuracy of 91.3%.

Corresponding Author:

Mutiara S. Simanjuntak,
Student Of Department of Electrical and Computer Engineering,
National Kaohsiung University Of Science and Tecnology,
Jiangong Campus. No 415 Jiangong Rd., Sanmin Dist., Kaohsiung City 807, Taiwan, ROC
Email: i110154114@nkust.edu.tw

1. INTRODUCTION

E-commerce is a container in the process of buying and selling goods and services online or the ability to transact online, including retail, online banking and shopping which involves transactions where buyers actually buy and shop[1]. Based on the results of the analysis that has been carried out, most of the consumers who have shopped at E-Commerce will provide a review of the products that have been purchased. Reviews and ratings are also very important to increase the frequency of customers, because reviews and ratings from customers can provide a more accurate and emotional assessment because they are given by fellow customers so that they have a higher trust value. Therefore, the rating of an online store requires special attention from the online store manager to

increase its customers[2]. However, customer confidence in the rating can be hampered due to inconsistency of answers such as reviews that have negative texts with positive rating values. The existence of different answers such as good ratings but bad reviews and vice versa can make other people confused[3].

Several previous studies, in conducting sentiment analysis, a method that supports classification is needed. It is stated that with Text Mining technology, it can solve complex problems[4]. States that for data mining techniques, appropriate methods and algorithms are needed to obtain appropriate results, so that at this time there are very many types of algorithms used in data mining techniques[5]. In conducted research on the application of the SVM Algorithm in analyzing sentiment on twitter data against the Corruption Eradication Commission of the Republic of Indonesia. The classification method used in this research is Support Vector Machine (SVM) and feature extraction using TF-IDF. Each occurrence of the word is labeled Positive, Negative, Neutral. Based on the test results, the application of the SVM method produces an accuracy value of 82% and produces a sentiment with a larger negative label with a total of 77%, a positive label of 8% and a neutral label of 25%[6]. The topic of this research is the analysis and classification of the existing comment data on Lazada. The process of classifying and identifying data is called Text Mining. The purpose of text mining (which is also referred to as data mining and text analysis) is to analyze textual documents from an unstructured form to a structured one so that it can be continued in the next stage of analysis both qualitatively and quantitatively[7]. The algorithm used in this research is SVM can be applied in a labeled data set, which will produce a series of input-output mappings labeled functions and feature details, SVM can also be used as a classification method[8]. Before processing the data, it must first go through the preprocessing stage to get better data and reduce noise. The data is taken from the Kaggle website in the form of Excel and then converted to a CSV file. For accuracy measurement, use confusion matrix for Precision, Recall, and F-measure assessment.

2. RESEARCH METHOD

2.1 Preprocessing Data

The dataset taken from the Kaggle site is still in the form of raw data so that the preprocessing stage is carried out to obtain clean data to facilitate the next stage and produce more accurate analysis results. The stages in Preprocessing are as follows:

1. Case folding is the process of converting all letters in the document into lowercase letters and removing all letters that are not alphabetical (az)[9]. Example: The item is good the item is good.
2. Cleaning is the process of cleaning documents from words that are less important in order to reduce noise in order to increase the accuracy of the classification process[10]. The words that will be removed from the comments are numbers, symbol characters, hashtags (#), changing the word alay to standard words and mentions (@username). Example: i love the same goooddddd I love the same good.
3. Stemming is looking for the basic word from the sentence and reducing it because it is a type of word that has the same meaning[10]. In this study using the library provided by Sastrawi.
4. Filtering is the process of saving word choices or deleting words[11]. The deleted words are words that have been categorized into Stopwords or general words that often appear in large numbers but have no meaning, including “and”, “or”, “to”, “di”, punctuation.
5. Tokenizing is the process of separating a full text string into a separate list of words[11]
6. Normalization is the weighting of sentences.
7. Term Frequency or TF is a word weighting process by adding up the words that appear in a document. While Inverse Document Frequency or IDF is the number of occurrences of a word in

all existing documents, and DF is obtained from the results of TF[12].

The TF formula can be seen in the following equation:

$$Tf_{t.d} = \quad if\ \mathrm{t}f_{t,d} > 0 \tag{1}$$

Otherwise

As for the IDF formula, it can be seen in the following equation:

$$IDF_t = log(\frac{N}{df_t}) \tag{2}$$

Where N is the number of all documents in the dataset and df t is the number of documents ontaining term t in it.

2.2 Analysis with Support Vector Machine

At this stage, an analysis of the classification method is carried out based on the data that has been processed using training data and the results of the analysis of the training data will be tested using test data. The flowchart of the SVM classification process is shown in Figure 2. The method used in SVM is a sequential method, which is a method that is useful for getting a hyperline, while the function of a hyperline is as a separator of two classes in the input space. The function of the sequential method is to speed up the iteration process. The following are the steps of the sequential method[13]:

1. Initialization is carried out on the parameters to be used, namely (lambda), (learning rate), C (complexity), (epsilon), and maximum iterations.
2. Initialize the value then calculate the matrix with equation 3.

$$D_{ij} = y_i\, y_j(\mathrm{K}(x_{1,} y_j) + \lambda^2) \tag{3}$$

3. Calculates equations (4), (5), and (6) to update the values of E and α.

a. $E_i = \sum_i^n \alpha_i\, ij$ (4)

b. $\delta_{\alpha i} = min\{max[\gamma(1 - E_i), -\alpha_i], C\text{-}\alpha_i\}$ (5)

c. $\alpha_i = \alpha_i + \delta_{\alpha i}$ (6)

4. Perform step 3 until maximum iteration or Max($/\delta\alpha/$)<ε.

After the above process is complete, it will be obtained and support vector. The next step is to calculate the value of b bias with equation 7.

$$\mathrm{b}=\text{-}\frac{1}{2}\left(\sum_{i=0}^{n} \alpha_i\, y_{i} K\left(x_{i,} x^{-}\right) + \sum_{i=0}^{n} \alpha_i\, y_{i} K\left(x_i, x^{+}\right)\right) \tag{7}$$

Sentiment analysis can be calculated using equation 8.

$$\mathrm{f(x)}=\sum_{i=0}^{n} \alpha_i\, y_{i} K(x, x_i) + b \tag{4}$$

2.3 Testing the classification results using the Confusion Matrix

Confusion Matrix is an important measure to evaluate the accuracy of the classification model. In the Confusion Matrix there is a binary classification type which only has 2 output classes which are shown in the following table:

Table 1. Binary Classification

Class	Classified Positive	Classified Negative
Positive	TP (True Positive)	TN (True Negative)
Negative	FP (False Negative)	FN (False Negative)

Confusion Matrix produces 3 outputs, namely:

$$\text{Recall} = \frac{TP}{FN+TP} X100\% \tag{9}$$

$$\text{Precission} = \frac{TP}{FP+TP} \mathrm{X}100\% \tag{10}$$

$$\text{Accuracy}=\frac{(TP+TN)}{(TP+TN+FP+FN)}X100\% \quad (11)$$

3. RESULT AND DISCUSSION

3.1 Research Dataset Analysis

To continue the analysis process, the dataset in the form of CSV is uploaded to Google Colab. The research dataset is called "datasetpenelitian.csv", then the data is imported and displayed on the system.

```
data[['Rating','Data Komentar']]
```

	Rating	Data Komentar
0	5	bagus mantap dah sesui pesanan
1	4	Bagus, sesuai foto
2	5	okkkkk mantaaaaaaapppp ... goood
3	4	bagus sesuai
4	1	baru 10 bulan layarnya dah bergaris
...	...	...
1040	5	semoga awet
1041	5	barang sesuai gambar dan berfungsi dengan baik...
1042	5	barang sudah diterima dan sesuai pesanan.
1043	5	pengirman ok banget, cepat.
1044	5	imut, blm sempat dicoba

1045 rows × 2 columns

Figure 1. Results of Research Data Display on the System

The labeling in this study was made into 3 categories, the Good category was given a label 1, the Neutral category was given a label 0, the Less good category was given a label -1. So the results of the system test based on the label given show that there are 696 comment data that are in the good category, 220 are in the neutral category, and 129 are in the less good category. After that, the system proceeds to the preprocessing stage and the classification process using SVM. See Figure 4.

```
def classes_def(x):
    if x == 5:
        return 1
    elif x == 4:
        return 0
    elif x == 3:
        return 0
    elif x == 2:
        return -1
    else:
        return -1

data['class']=data['Rating'].apply(lambda x:classes_def(x))

print("Bagus: ", data[data['class'] == 1].shape)
print("Netral: ", data[data['class'] == 0].shape)
print("Kurang Bagus: ", data[data['class'] == -1].shape)
```

```
Bagus:  (696, 4)
Netral:  (220, 4)
Kurang Bagus:  (129, 4)
```

Figure 2. Data Display Results That Have Gone Through The Labeling Stage

To support the preprocessing process , there are two CSV files that are imported to the system, namely files to remove words/term stopwords and change words that are included in the "alay" dictionary, while the purpose of these two supporting files is to assist the data cleaning process. which is more accurate. See Figure 5.

```
[77] #2.1. Import file CSV kumpulan stopword
     id_stopword_dict = pd.read_csv('stopwordbahasa.csv')
     id_stopword_dict = id_stopword_dict.rename(columns={0: "stopword"})

     print("Shape: ", id_stopword_dict.shape)
     id_stopword_dict.head()

     Shape:  (758, 1)
```

	stopword
0	ada
1	adalah
2	adanya
3	adapun
4	agak

Figure 3. Display Stopword

```
#2.2. Import file CSV kumpulan kamus alay
alay_dict = pd.read_csv('new_kamusalay.csv')
alay_dict = alay_dict.rename(columns={0: 'original', 1: 'replacement'})

[73] print("Shape: ", alay_dict.shape)
     alay_dict.head(15)

     Shape:  (15169, 2)
```

	original	replacement
0	anakjakartaasikasik	anak jakarta asyik asyik
1	pakcikdahtua	pak cik sudah tua
2	pakcikmudalagi	pak cik muda lagi
3	t3tapjokowi	tetap jokowi
4	3x	tiga kali
5	aamiin	amin
6	aamiinn	amin
7	aamin	amin

Figure 4. Display of Alay Dictionary

3.2 Preprocessing Results

3.2.1 Text Preprocessing

The results of the text Preprocessing data process :

a. Case Folding

Based on the preprocessing method , the words of each comment will be changed to lowercase as shown in the following figure:

	Data Komentar
0	bagus mantap dah sesui pesanan
1	bagus, sesuai foto
2	okkkkk mantaaaaaaapppp ... goood
3	bagus sesuai
4	baru 10 bulan layarnya dah bergaris
...	...
1040	semoga awet
1041	barang sesuai gambar dan berfungsi dengan baik...
1042	barang sudah diterima dan sesuai pesanan.
1043	pengirman ok banget, cepat.
1044	imut, blm sempat dicoba

1045 rows × 1 columns

Figure 5. Results of Preprocessing with Casefolding

b. Cleaning

Based on the Preprocessing method , the results of the Cleaning process show that words that are considered less important have been removed automatically at this stage, as shown in the following figure:

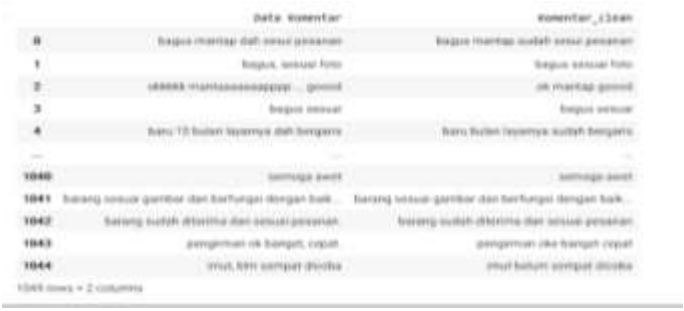

Figure 6. Results of Preprocessing with Cleaning

c. Stemming

Based on the Preprocessing method , the results of the Stemming process are obtained, namely removing affixes and retaining basic words as shown in the following figure:

Figure 7. Preprocessing Results with Stemming

d. Filtering

Based on the Preprocessing method , the results of the Filtering process are obtained, namely common words have been removed to reduce noise as shown in the following figure:

Figure 8. Results of Preprocessing with Filtering

3.3 Normalization of TF-IDF

Based on equations (1) and (2) with the weighting stage for each word (term), the results of the TF-IDF normalization are as follows:

```
array([[0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       ...,
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0]])
```

Figure 9. Results of Preprocessing with TF-IDF

3.4 Data Visualization

In this study, there are two data visualization techniques used, namely as follows:

a. Histogram.

Histogram serves to display the relationship between the length of the character and the frequency of data in a document [18]. Below shows the histogram of the comment data visualization results.

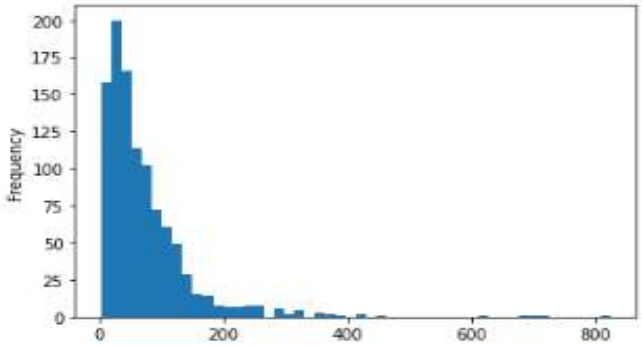

Figure 10. Histogram of Research Dataset

b. Word cloud. Wordcloud serves to visualize documents in the form of text so that they have an attractive appearance. Below are the results of visualization of comment documents from consumers who shop at Lazada, the words that are large are the words that appear most often [19].

Figure 11. Wordcloud Research Dataset

3.5. Sharing training data and test data (split data)

```
[25] from sklearn.model_selection import train_test_split
     # pemisahan train dan test
     X_train, X_test, Y_train, Y_test = train_test_split(X,Y,test_size=0.1, random_state=100)
```

Figure 12. Split test data and training data

3.6 Classification with SVM Algorithm

As for the prediction results of the SVM classification from the data that has been divided into training data and test data. The percentage value of the accuracy level is generated from the comment data that has been labeled with the weighted value of each calculated document. The time in the process of completing the SVM Analysis program is 11.575 seconds. The following is a picture of the results of the SVM classification :

Figure 13. SVM Klasifikasi Classification Results

The next step is to calculate the level of accuracy of the SVM algorithm on research training data. The level of accuracy of the classification of the SVM algorithm on the training data is 91.38%.

```
# Prediksi Data Training SVM
predicted = model.predict(X_train)
# membuat classification report (precision, recall, f1-score,support)
target_names = ['Kurang Bagus', 'Netral', 'Bagus']
report = classification_report(Y_train,predicted,target_names=target_names)
# "precision: tp/(tp+fp)"
# "recall: tp/(tp+fn)"
# "f1-score: (2xprecisionxrecall)/(precision+recall)"
#split classification report
print(report)
```

```
              precision    recall  f1-score   support

Kurang Bagus       0.88      0.91      0.89       120
      Netral       0.81      0.76      0.79       199
      Bagus        0.92      0.93      0.93       621

    accuracy                           0.89       940
   macro avg       0.87      0.87      0.87       940
weighted avg       0.89      0.89      0.89       940
```

Figure 14. SVM Classification Results Report based on training data

Next, calculate the level of accuracy of the SVM algorithm on the research test data. The value of the classification accuracy of the SVM algorithm from the test data is 71.42%.

```
# Prediksi Data Testing SVM
predicted = model.predict(X_test)
# membuat classification report (precision, recall, f1-score,support)
target_names = ['Kurang Bagus', 'Netral', 'Bagus']
report = classification_report(Y_test,predicted,target_names=target_names)
# "precision: tp/(tp+fp)"
# "recall: tp/(tp+fn)"
# "f1-score: (2xprecisionxrecall)/(precision+recall)"
#split classification report
print(report)
```

```
              precision    recall  f1-score   support

Kurang Bagus       0.56      0.42      0.48        12
      Netral       0.44      0.33      0.38        21
       Bagus       0.79      0.88      0.83        72

    accuracy                           0.71       105
   macro avg       0.59      0.54      0.56       105
weighted avg       0.69      0.71      0.70       105
```

Figure 15. SVM Classification Result Report based on test data

3.7 Analysis of Classification Results with Confusion Matrix

From the application of equation (9), the Confusion Matrix results from the SVM analysis for the Good category have a recall value of 79%. With equation (10), SVM analysis has a precision level of 88%, and in equation (11), SVM analysis has an accuracy rate of 71%. After testing the classification based on commentary and rating data simultaneously using SVM for the less good and neutral categories, the results are lower than the good categories.

Following are the results of the Confusion Matrix SVM analysis:

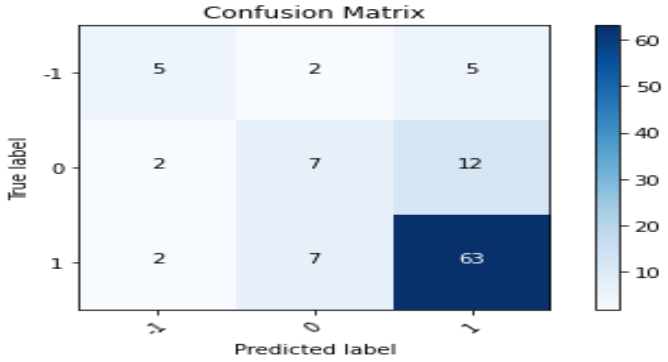

Figure 16. Confusion Matrix Diagram

3. CONCLUSION

The dataset used in this study is 1044 comment data along with ratings from the Lazada platform. Then the comment data along with the rating are labeled with good, neutral, and bad categories, and classified using the SVM algorithm and the level of accuracy using the Conffusion Matrix. However, in this study only the classification process for datasets that have been inputted previously, not for datasets that have just been inputted. SVM classification resulted in training data classification accuracy of 91.3% and test data classification accuracy of 71.4% with a compiler time of 11,575 seconds. For further research, do a classification using more data and higher preprocessing techniques so as to get clean data and more accurate classification results.

ACKNOWLEDGEMENTS

This research can be carried out properly thanks to the help of various parties, for that the researchers would like to thank for the support and good cooperation from various parties.

REFERENCES

[1] T. H. D. Chaffey, D. Edmundson-Bird, Digital Business and E-Commerce Management, 7th Edition. UK: Pearson UK, 2019. [Online]. Available: https://www.pearson.com/uk/educators/higher-education-educators/program/Chaffey-Digital-Business-and-E-Commerce-Management-7th-Edition/PGM2542799.html

[2] D. Pujiwidodo, "The influence of online customer reviews and ratings on trust and purchase interest in online marketplaces in Indonesia," vol. III, no. 2, p. 2016, 2016.

[3] Regina Dwi Amelia, M. Michael, and R. Mulyandi, "Online Consumer Review Analysis of Purchase Decisions in Beauty E-Commerce," J. Indones. Sos. Teknol., vol. 2, no. 2, pp. 274–280, 2021, doi: 10.36418/jist.v2i2.80.

[4] D. Maulina and R. Sagara, "Indonesia. Sauce. Technol. , vol. 2, no. 2, pp. 274–280, 2021. [4] D. Maulina and R. Sagara, "Classification of hoax articles using linear support vector machines with weighting term frequency–Inverse document frequency," J. Mantik Penusa, vol. 2, no. 1, pp. 35–40, 2018.

[5] N. M. Norwawi, "Recognition decision-making model using temporal data mining technique".

[6] D. Darwis, E. S. Pratiwi, and A. F. O. Pasaribu, "Application of the Svm Algorithm for Sentiment Analysis on Twitter Data of the Corruption Eradication Commission of the Republic of Indonesia," Edutic - Sci. J. Informatics Educ., vol. 7, no. 1, pp. 1–11, 2020, doi: 10.21107/edutic.v7i1.8779.

[7] M. P. Bach, Ž. Krstič, S. Seljan, and L. Turulja, "Text mining for big data analysis in financial sector: A literature review," Sustain., vol. 11, no. 5, 2019, doi: 10.3390/su11051277.

[8] A. K. Gupta, V. Singh, P. Mathur, and C. M. Travieso-Gonzalez, "Prediction of COVID-19 pandemic measuring criteria using support vector machine, prophet and linear regression models in Indian scenario," J. Interdiscip. Math., vol. 24, no. 1, pp. 89–108, 2021, doi: 10.1080/09720502.2020.1833458.

[9] M. S. Simanjuntak, "Decision Support System For Admission Of New Employees With Fuzzy Madm Model Weighted Product At Pt. Super Andalas Stell," Univ. POTENSI UTAMA, 2020.

[10] B. Gupta, M. Negi, K. Vishwakarma, G. Rawat, and P. Badhani, "Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python," Int. J. Comput. Appl., vol. 165, no. 9, pp. 29–34, 2017, doi: 10.5120/ijca2017914022.

[11] M. S. Simanjuntak and J. Panjaitan, "Information Retrieval System Using K- Nearest Neighbour In Journal Classification," vol. 1, no. 2, pp. 1–8, 2021.

[12] M. S. Simanjuntak, "Jurnal Mantik Jurnal Mantik," Act. Act. Funct. Multilayer Perceptron - Based Card. Abnorm., vol. 3, no. 2, pp. 10–19, 2019, [Online]. Available: http://iocscience.org/ejournal/index.php/mantik/article/view/747/530

[13] R. Rosnelly, D. Hartama, M. Sadikin, C. Lubis, M. Simanjuntak, and S. Kosasi, "The Similarity of Essay Examination Results using Preprocessing Text Mining with Cosine Similarity and Nazief-Adriani Algorithms," Turkish J. Comput. Math. Educ., vol. 12, pp. 1415–1422, Apr. 2021.