JINAV: Journal of Information and Visualization *ISSN: 2746-1440 (online)*
Vol. 3 No. 1 (2022)
https://doi.org/10.35877/454RI.jinav1389

Text Classification on Sentiment Analysis of Marketplace SHOPEE Reviews On Twitter Using K-Nearest Neighbor (KNN) Method

Zulkifli Rais*, Rezky Novita Said, & Ruliana

Department of Statistics, Universitas Negeri Makassar, Makassar, Indonesia

Abstract

This research aims to know the description and result of the classification sentiment analysis by Twitter users about Shopee. The method used in this research is K-Nearest Neighbor. K-Nearest Neighbor is a method that identifies groups or classifications based on the closest k of test data (training data). The most relative distance is calculated using the Euclidean distance. The data in this research were obtained from the Twitter API which used data on July 13, 2021, and the data according to the study were 150 tweets. Based on the results of the preprocessing text, there are 10 words that appear most often conveyed by Twitter users, and these opinions are related to the features provided by Shopee. The results obtained from this research are the highest level of text classification accuracy is 90% in training data and testing data comparison 80%: 20%

Keywords: Sentiment analysis; K-Nearest Neighbor; Twitter; Shopee.

1. Introduction

Technological developments are very beneficial for all people. With the development of technology, it indirectly causes the development of transactions, one of which is the purchase of goods using online shopping media (Mahayani et al., 2020). Some terms in online shopping media are e-commerce, online shop and marketplace (Sakhrin, 2018). Several types of marketplaces that are popular in Indonesia are Lazada, Tokopedia, Bukalapak and Shopee (Ardianti & Widiartanto, 2019).

Every purchase through the marketplace will begin with some consumer considerations by looking at the reviews of consumers who have bought before (Muktafin et al., 2020). Reviews play an important role for consumers because they provide a reference for them in determining the decisions to be taken as well as obtaining opinions from the personal experiences of people who have previously purchased goods or experienced the services offered (Wang & Wang, 2020). To make it easier for consumers to see previous consumer reviews, one method that can be used to classify the types of reviews given by consumers is text classification.

Text classification is a process in determining the similarity of texts according to predetermined categories (Chakraborty et al., 2014). Several statistical methods that can be used in text classification include K-Nearest Neighbor (KNN) and Naïve Bayes (Kolluri & Razia, 2020). According to research results by (Pertiwi, 2019) the accuracy of the K-Nearest Neighbor (KNN) method is 90.76%. In addition, according to research results by (Hamdana et al., 2020) the accuracy of the K-Nearest Neighbor (KNN) method is 88%.

Sentiment analysis is a part of text classification that aims to extract subjective information from texts in the form of natural language, opinions and sentiments which can then become structured and actionable information in decision

* Corresponding author.
E-mail address: zulkifli.rais89@unm.ac.id

making (Pozzi et al., 2017). In the business world, sentiment analysis is usually used in analyzing customer reviews or comments on the services or products provided (Fitri et al., 2019). Reviews from customers are very meaningful in the business world because they are needed in determining the policies to be taken (Liu, 2015). Reviews in the form of text can be divided into 3 classifications, namely positive, negative and neutral classifications. Currently, reviews or opinions can be channeled through social media. Some types of social media that are commonly used by people are Facebook, Instagram and Twitter. Twitter is one of the popular social media used by the public to channel opinions, opinions or reviews (Wicaksono, 2020).

Based on the circumstances, this study conducts Classify in text classification on Sentiment Analysis of Shopee marketplace reviews on Twitter Social Media using the K-Nearest Neighbor (KNN) Method..

2. Literature Review

Classification is the process of processing data to be grouped into certain classes. Classification performs modeling derived from training data which is then used to classify data on testing data (Utomo & Mesran, 2020). One of the classification methods commonly used is K-Nearest Neighbor (KNN).K nearest neighbor is a method that can identify groups or classifications (Riadhi et al., 2020). Classification identification aims to determine the classification of the tested data based on the closest k on the test data (training data). The nearest k data can be obtained by calculating the Euclidean distance interval (Zakaria et al., 2020). The proximity through the interval can be calculated using the Euclidean distance which is written in the equation 1.

$$d_{(i,j)} = \sqrt{\left(\left|x_{i1} - x_{j1}\right|^2 + \left|x_{i2} - x_{j2}\right|^2 + \cdots + \left|x_{ip} - x_{jp}\right|^2\right)} \qquad (1)$$

where :

$d_{(i,j)}$ = distance between document *i* and document *j*
$x_{i(n)}$ = the nth word in the *i* document
$x_{j(n)}$ = the nth word in the *j* document

The stages in implementing the K-Nearest Neighbor (KNN) method as follows (Irfan, 2018):

a. Entering training data and testing data
b. Calculating the similarity of training data and testing data
c. Sorts the results of the similarity of data from the largest to the smallest
d. Determine the k training data that have the closest resemblance to the testing data
e. Checking data label *k*
f. Define label
g. Counting accuracy

Counting accuracy aims to determine how much accuracy the data is categorized based on the confusion matrix (Suniantara dkk., 2020). confusion matrix serves to provide an assessment of the classification ability based on true and false (Rahayuningsih, 2019). Table 1 will describe in general the confusion matrix.

Table 1. Confusion Matrix

	Positive	**Negative**
Actual Positive	True Positif (TP)	False negative (FN)
Actual Negative	False Positif (FP)	True Negative (TN)

3. Research Method

The steps taken in this study are as follows:

1) Manage twitter data open source licensing as data source of this research
2) Collecting data in form of a collection tweets containing the keyword "@ShopeeID" obtained from twitter by utilizing the API Twitter
3) Performing data pre-processing which consists of 4 steps, namely:
   a. Case folding, the process of changing capital letters to lowercase letters
   b. Punctuation removal, the process of removing symbols and punctuation marks
   c. Tokenization, the process of converting sentences into words
   d. Stop word removal, the process of removing conjunctions, third person pronouns, characters and numbers
4) Make a list of words that are included in the positive and negative classification to determine the score that is used as the basis for determining the classification in the manual tagging method. The glossary is obtained from the github.com page
5) Determine the classification of documents using the manual tagging method. Documents are classified as positive if the document scores > 0, negative classification if the document score 0, and neutral classification if the score = 0.6. Do the word weighting using the term Frequency-Inverse Documentary Frequency (TF-IDF)
6) Do the word weighting using the term Frequency-Inverse Documentary Frequency (TF-IDF)
7) Divide the data into testing data and training data with a ratio of 70%: 30% and 80%:20%. The training data is used to test the KNN algorithm that has been created, while the testing data is used to determine the accuracy of the classification results (Claudy et al., 2018)
8) Determine the similarity of the sentences of testing data and training data using the Euclidean distance according to equation 1
9) Classification of review sentiments using training data and testing data by determining the value of k
10) Analysed the classification results
11) Calculating method accuracy, using confusion matrix
12) Interpretation of analysis results
13) Drawing conclusions

4. Results and Discussion

4.1. Results

The amount of data collected is 235 data. The keyword used in this research is @ShopeeCare as a Shopee's customer service account on Twitter. The form of the data that has been collected can be seen in Figure 1.

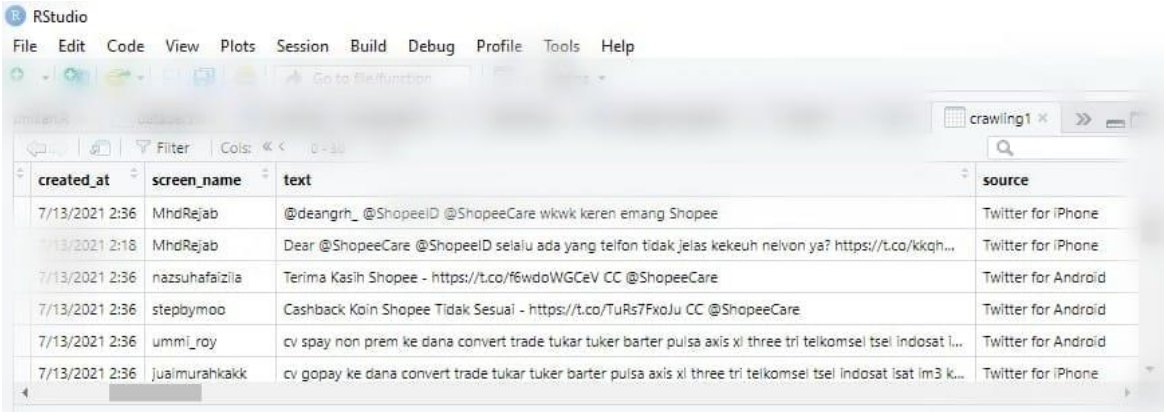

Figure 1. Data Collection Results

Tweet data that have been collected and then in preprocessing includes the process of case folding, punctuation removed, stopword removal and tokenization. The results before and after preprocessing can be seen in Table 2.

Tabel 2. Result of preprocessing text

Before Preprocessing Text	After Preprocessing Text
@deangrh_ @ShopeeID @ShopeeCare wkwk keren memang Shopee	*keren memang*
Dear @ShopeeCare @ShopeeID selalu ada yang telfon tidak jelas kekeuh nelvon ya? https://t.co/kkqhWkb2Y5	*selalu ada telfon tidak jelas*
Terima Kasih Shopee - https://t.co/f6wdoWGCeV CC @ShopeeCare	*terima kasih*
⋮	⋮
tarik tunai dr shopee pay ada biaya admin ya allah menangiz	*tarik tunai ada biaya admin*

After the text preprocessing process is carried out, the data will be categorized using the manual tagging method. The basis for determining the final score is the difference between the number of positive words and negative words. The results of determining the category of Manual Tagging based on scores can be seen in table 3.

Tabel 3 Result of *Manual Tagging*

Text	Score	Category
keren memang	1	Positif
selalu ada telfon tidak jelas	-1	Negatif
terima kasih	2	Positif
⋮	⋮	⋮
tarik tunai ada biaya admin	-1	Negatif

After preprocessing the text, the results of the process will calculate the weight of each word using the term frequency inverse documentary frequency (TF-IDF) word weighting method. The results of word weighting can be seen in Table 4.

Table 4. Result of Term Weighting using TF-IDF

Term	*keren memang*	*selalu ada telfon tidak jelas*	*terima kasih*	…	*tarik tunai ada biaya admin*
Account	0	0	0	⋯	0
Ada	0	4.0690	0	⋯	4.0690
Admin	0	0	0	⋯	4.0690
Ajukan	0	0	0	⋯	0
Akan	0	0	0	⋯	0
Akun	0	0	0	⋯	0
⋮	⋮	⋮	⋮	⋯	⋮
Yakin	0	0	0	⋯	0

The data used in testing the classification using the K Nearest Neighbor method are:

1. Distribution of training and testing data 70% : 30%
2. Distribution of training and testing data 80% : 20%

Based on the classification process using the K Nearest Neighbor method which was tested on 10 k values, Table 5 will show the accuracy based on the results of the k value test.

Table 5. Accuracy Comparison *k* values

***k* values**	**Data Training and Data Testing 70 % : 30%**	**Data Training and Data Testing 80 % : 20%**
1	82%	90%
2	72%	70%
3	64%	60%
4	60%	60%
5	64%	56.67%
6	60%	60%
7	56%	56.67%
8	54%	53.33%
9	58%	56.67%
10	54%	56.67%

4.2. Discussion

Based on Figure 2, it can be seen that the percentage of negative opinions is higher than that of positive and neutral opinions. Of the 150 data that have gone through preprocessing text and manual tagging, 51% have negative opinions, 35% positive opinions and 14% neutral opinions. Comparison of the results of the accuracy of k nearest neighbor in 2 types of distribution of training data and testing data can be seen in Figure 3.

It can be seen that the graph tends to decrease as the value of k increases. It shows that the greater the value of k, the smaller the accuracy produced, for example, the value of k = 1 each has an accuracy of 82% and 90%. This can happen because the testing data only takes into account 1 nearest neighbor, it is different if the value of k is added then the testing data will take into account the various neighbor values. This is in line with research conducted by (Mentari et al., 2018) which showed that of the 9 tested k values, the k value = 1 had the highest accuracy, which was 93.63% and the accuracy would decrease when the k value was added.

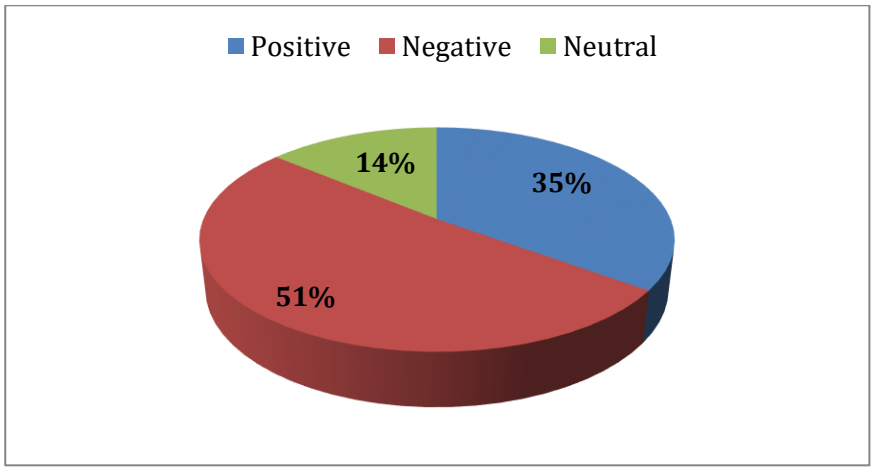

Figure 2 Percentage of users opinions on Twitter

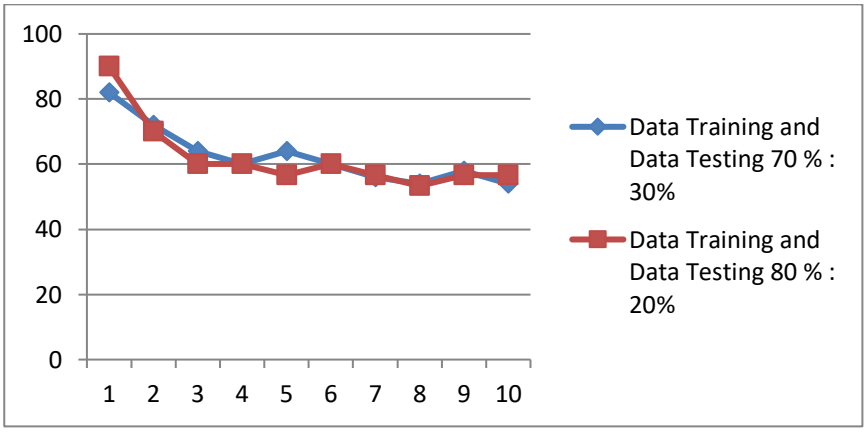

Figure 3. Accuracy results of KNN Method

5. Conclusion

Based on the text classification in the Shopee marketplace review sentiment analysis on Twitter, it can be concluded as follows:

a. The results of preprocessing text as much as 150 data, there are 10 words that appear most often conveyed by Twitter users where the majority of these opinions are related to features provided by shopee such as promos, free shipping (postage) to relate to Shopee digital money, namely shopeepay.

b. Based on the results of the classification using the KNN method, the accuracy obtained by testing 10 k values in the distribution of training data and testing data is 70%: 30% having an accuracy of 83% and for the distribution of training data and testing data 80%: 20% have an accuracy of 90%.

References

Ardianti, A. N., & Widiartanto, M. A. (2019). Pengaruh Online Customer Review dan Online Customer Rating terhadap Keputusan Pembelian melalui Marketplace Shopee . *Jurnal Ilmu Administrasi Bisnis*, 1–11.

Chakraborty, G., Pagolu, M., & Garla, S. (2014). *Text Mining and Analysis Practical Methods, Examples, and Case Studies Using SAS*. SAS Institute.

Claudy, Y. I., Perdana, R. S., & Fauzi, M. A. (2018). Klasifikasi Dokumen Twitter Untuk Mengetahui Karakter Calon Karyawan Menggunakan Algoritme K-Nearest Neighbor (knn). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, *2*(8), 2761–2765. https://www.researchgate.net/publication/322959490

Fitri, V. A., Andreswari, R., & Hasibuan, M. A. (2019). Sentiment analysis of social media Twitter with case of Anti-LGBT campaign in Indonesia using Naïve Bayes, decision tree, and random forest algorithm. *Procedia Computer Science*, *161*, 765–772.

Hamdana, E. N., Balya, M., & Alfahmi, I. (2020). *Pengembangan Sistem Analisis Sentimen Berbasis Java Pada Data Twitter Terhadap Omnibus Law Menggunakan Algoritma Naïve Bayes dan K- Nearst Neighbor (K-NN)*. 79–84.

Irfan, M. R. (2018). *Analisis Sentimen Kurikulum 2013 Pada Twitter Menggunakan Ensemble Feature Dan Metode K Nearest Neighbor*. Universitas Brawijaya.

Kolluri, J., & Razia, S. (2020). Text classification using Naïve Bayes classifier. *Materials Today: Proceedings*.

Liu, B. (2015). sentiment analysis. In *Cambridge University Press*. https://doi.org/10.14569/ijacsa.2018.090981

Mahayani, I., Agushinta, D., Muhammad, R., Supriyadi, E., Ilmu, F., Informasi, T., Gunadarma, U., Margonda, J., No, R., & Barat, J. (2020). *ShopeePayLater Pada Aplikasi Belanja Online (Shopee) Menggunakan Metode Lexicon Based dan Naïve Bayes Classi er Pendahuluan Naïve Bayes Classi- Analisis Sentimen dan Klasi-*. *19*, 545–558.

Mentari, N. D., Fauzi, M. A., & Muflikhah, L. (2018). Analisis Sentimen Kurikulum 2013 Pada Sosial Media Twitter Menggunakan Metode K-Nearest Neighbor dan Feature Selection Query Expansion Ranking. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, *2*(8), 2739–2743.

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.

Pertiwi, M. W. (2019). Analisis Sentimen Opini Publik Mengenai Sarana Dan Transportasi Mudik Tahun 2019 Pada Twitter Menggunakan Algoritma Naïve Bayes, Neural Network, KNN dan SVM. *Inti Nusa Mandiri*, *14*(1), 27–32.

Pozzi, F. A., Fersini, E., Messina, E., & Liu, B. (2017). Challenges of Sentiment Analysis in Social Networks: An Overview. In *Sentiment Analysis in Social Networks* (hal. 1–11). Elsevier Inc.

Rahayuningsih, P. A. (2019). Komparasi Algoritma Klasifikasi Data Mining untuk Memprediksi Tingkat Kematian Dini Kanker dengan Dataset Early Death Cancer. *Jurnal Teknik Informatika Kaputama (JTIK)*, *3*(2).

Riadhi, A. R., Aidid, M. K., & Ahmar, A. S. (2020). Analisis Penyebaran Hunian dengan Menggunakan Metode Nearest Neighbor Analysis. *VARIANSI: Journal of Statistics and Its application on Teaching and Research*, *2*(1), 46. https://doi.org/10.35580/variansiunm12901

Sakhrin, R. K. M. A. (2018). *Study Reception Terhadap Fenomena Situs Belanja Online Berbasis E-Commerce*. Universitas Pembangunan Nasional Veteran Jakarta.

Suniantara, I. K. P., Suwardika, G., & Soraya, S. (2020). Peningkatan Akurasi Klasifikasi Ketidaktepatan Waktu Kelulusan Mahasiswa Menggunakan Metode Boosting Neural Network. *Jurnal Varian*, *3*(2), 95–102. https://doi.org/10.30812/varian.v3i2.651

Utomo, D. P., & Mesran, M. (2020). Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung. *Jurnal Media Informatika Budidarma*, *4*(2), 437. https://doi.org/10.30865/mib.v4i2.2080

Wang, H., & Wang, Y. (2020). A Review of Online Product Reviews. *Journal of Service Science and Management*, *13*(01), 88–96.

Wicaksono, W. A. (2020). (2020). *Pengaruh Persepsi Risiko, Online Customer Review Dan Kepercayaan Terhadap Keputusan Pembelian Secara Online Di Shopee (Studi Kasus Pada Mahasiswa Fakultas Ekonomika Dan Bisnis Universitas Stikubank Semarang)*. Universitas Stikubank.

Zakaria, L., Ebeid, H. M., Dahshan, S., & Tolba, M. F. (2020). Analysis of Classification Methods for Gene Expression Data. In *Advances in Intelligent Systems and Computing* (Vol. 921). Springer International Publishing