Sentiment Analysis on Tokopedia Product Online Reviews Using Random Forest Method

Stephenie[1*], *Budi* Warsito[2], *Alan* Prahutama[3]

[1]Bachelor Program of Statistics, Faculty of Science and Mathematics, Diponegoro University, Semarang – Indonesia
[2,3]Department of Statistics, Faculty of Science and Mathematics, Diponegoro University, Semarang – Indonesia

Abstract. Tokopedia is one of the most popular e-commerce sites in Indonesia that offers consumer products from various categories. In each product section, a review feature is offered. This review feature became essential in evaluating the sellers and become one consideration for customers in making purchase consideration. Sentiment analysis of Tokopedia product reviews may provide the opportunity to look on how Tokopedia customers respond to product quality and sellers' hospitality. In evaluating the model, the reviews were grouped as: "positive sentiment" and "negative sentiment" using the Random Forest method and 10-fold cross-validation. Data labelling was carried out automatically by calculating the sentiment score using Lexicon-Based. Visualization of the labelling results was then done using a bar graph and a word cloud on each class of sentiment in order to look up for information that is considered important and most discussed. The test results showed that the accuracy of the Random Forest Method with parameter mtry = 73 and ntree = 50 is 97.38% which leads to the conclusion that the Random Forest Method could well predict the product reviews of Tokopedia. The greater the accuracy, the better performance of the classification model.

Keywords. Product Reviews, Tokopedia, Random Forest, Text mining, Word Cloud

1 INTRODUCTION

Tokopedia is one of the most popular e-commerce sites in Indonesia. Based on the research by CupoNation, Tokopedia has become the top list of the most popular e-commerce sites in Indonesia, with number of visits were as high as 1.2 billion [1]. Tokopedia offers various type of products and each product has a review feature. Product reviews is often used as an effective and efficient method to discover some information and to summarize the product quality. Upon the basis of previous statements, researcher wants to carry out sentiment analysis of Tokopedia review product in order to know the opinions of buyers and for the evaluation of sellers.

* Corresponding author: stephenie98@gmail.com

In this study, the reviews were grouped into two classes, which is positive and negative sentiment. Random Forest method was used because this method yields higher accuracy than the logistic regression and the decision trees, which will generate a large number of classification trees and will be aggregated in order to obtain the highest accuracy [2]. The researcher would also visualize each class of sentiment using word cloud in order to look up for the most discussed information on each class. Word cloud is a textual data visualization technique which presents the key words of data that generally consists of the most frequent words of the analyzed data.

2 LITERATURE REVIEW

2.1 Descriptive Statistics and Sentiment Analysis

Descriptive Statistics is one of the basic statistical methods used to analyze data by collecting and presenting the data in the form of narrative text, table, and diagram for the purpose of useful information from the visualization. Generally, data used in this analysis are from survey, census, or other observations that is not well organized and systematic.

Sentiment Analysis is a method used to process diversified opinions given by consumers or experts on products, services, or even institutions through various type of media [3]. In general, the purpose of sentiment analysis is to solve two classes classification problems which is positive and negative.

2.2 Text Mining

Text mining is knowledge-intensive process in which users interact with document collection over time by using a suite of analysis tools, one of which is categorization [4]. Text mining aims to transform unstructured textual data into structured data in order to obtain new information. There are two phase in text mining, which consist of data pre-processing and feature selection. Data pre-processing is the earliest phase in the preparation of textual data which consist of spelling normalization, case folding, tokenizing, filtering, and stemming. Feature selection is an advanced phase to reduce word dimensions with the relevant terms that actually represents the contents of document.

2.3 Data Labelling

Data labelling is carried out automatically by calculating the sentiment score using Lexicon-Based. The approach used to describe the sentiment class is by subtracting the total number of positive words and the total number of negative words. A sentence with score > 0, will be classified as positive class, a sentence with score = 0, will be classified as neutral class, while a sentence with score < 0, will be classified as negative class [5].

2.4 TF-IDF Weighting

The formula for the TF-IDF weighting can be described as following [6]:

$$W_{(d,t)} = TF_{(d,t)} \times log \frac{N}{df_{(t)}} \quad (1)$$

where:

$W_{(d,t)}$ = TF-IDF weighting of term t in document d.

$TF_{(d,t)}$ = The frequency of term t occurs in document d

N = The total number of documents in collection.
$df_{(t)}$ = The number of documents that contain the term t .

2.5 Word Cloud

Word cloud is textual data visualization that shows the conclusion of the data based on the frequency of each words. It is a cloud-shaped figure that consists of the most common terms in the textual analyzed data. The bigger the word in word cloud, the higher the intensity of the words. Word cloud is used to present the key words of the data in an interesting way, however it has a disadvantage of not being able to show the frequency of words in the analyzed text.

2.6 Random Forest

Random Forest Algorithm is the advancement of Classification and Regression Tree (CART) method with the implementation of bootstrap aggregating (bagging) and random feature selection [7]. Procedure of random forest algorithm on the data of n observations and p predictor [7,8]:

1. Random samples of size n are drawn with the possibility of obtaining the same data (with replacement). This phase is called bootstrap.
2. Using the bootstrap samples, the tree is grown until the maximum size is reached, which is done without pruning. At each node, the random feature selection is used to determine the split, which m number of variables randomly sampled as candidates at each split must be m << p, at which point, the best node will be chosen based on m number of variables available for splitting.
3. Repeat stage 1 and 2 for k times to generate a forest that consists of k trees.

Breiman and Cutler [8] suggests to observe the error OOB when $m = \left(\frac{1}{2}\left|\sqrt{p}\right|, \left|\sqrt{p}\right|, 2\left|\sqrt{p}\right|\right)$ where p is the total variable and the number of k is small, then m with the smallest error OOB will be chosen.

In order to determine the split used as root node/node, Gini Index is used in Random Forest method. The formula of Gini index can be described as following:

$$\boldsymbol{Gini = 1 - \sum_{i=1}^{k} p_i^2} \quad (2)$$

where:
p_i: probability of an attribute being classified to class i.
k: total number of attributes being classified to a particular class.

The number of k suggested to apply in bagging is k = 50 which will provide satisfied results for classification [7].

2.7 K-Fold Cross Validation

K-Fold Cross Validation aims to test the average accuracy of machine learning models by doing repetition with randomize the input attributes. This method begins by randomly splitting the data set into n folds. In cross validation, the dataset is split into n partition of approximately equal size D_1, D_2, D_3,..D_n, and the holdout method is repeated n times. In the i iteration, the D_i partition will be the testing set and the rest will be the training set [9].

2.8 Evaluation Model using *Confusion Matrix*

Evaluation is needed to measure the classifier performance. Performance of a classifier must be measured in order to know the accuracy of the predictions. For a binary classification problem can be represented with 2 x 2 matrix as shown in the table below:

Table 1. Confusion matrix

		Actual	
	Class	***Positive***	***Negative***
Predicted	***Positive***	*True Positive* (TP)	*False Positive* (FP)
	Negative	*False Negative* (FN)	*True Negative* (TN)

One of the model performance parameter is accuracy. Accuracy is the ratio of correct predictions to total number of predictions made. The formula can be expressed as following [10]:

$$Accuracy = \frac{TP+TN}{(TP+FP+TN+FN)} \quad (3)$$

3 RESEARCH METHOD

Data used in this research is Tokopedia product reviews in Indonesian language with the total number of 40,607 reviews. The data was obtained from the Tokopedia website database. Variables used in this method consist of rating and review. The rating defines the score of the products, while the review defines the content description of products. Software used in this data analysis were Anaconda Navigator 1.9.12 (Python), R 4.0.0, and Microsoft Excel 2010. The method used in this analysis were descriptive statistics which show the universal illustration about Tokopedia product reviews. The next phase is data pre-processing, followed by data labelling and TF-IDF weighting. After being labelled, word cloud was created in order to visualize the key words of each sentiment class. The last phase is data classification using random forest method in order to find the accuracy of the model and method used to evaluate the model were 10-fold cross validation and confusion matrix. Data pre-processing, TF-IDF weighting, and random forest classification were done by Python, while the data labelling and visualization through word cloud were done by R 4.0.0.

4 RESULTS AND DISCUSSION

4.1 Descriptive Statistics

The following pie chart below illustrates the reviews submitted based on product categories.

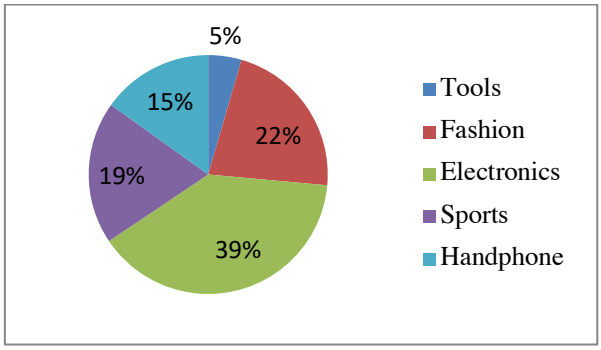

Figure 1. The reviews submitted based on product categories

Figure 1 shows the distribution of the reviews submitted on products from five categories, namely tools, fashion, electronics, sports, and handphone. Electronics has the highest percentage of 39%, which equal to 15,897 out of 40,607 reviews. It is followed by the fashion category with the proportion of 22%, and the minority category is tools which has accounted for 5% of the reviews. From the pie chart, it is clear that the majority of the customers has purchased in electronics category in Tokopedia.

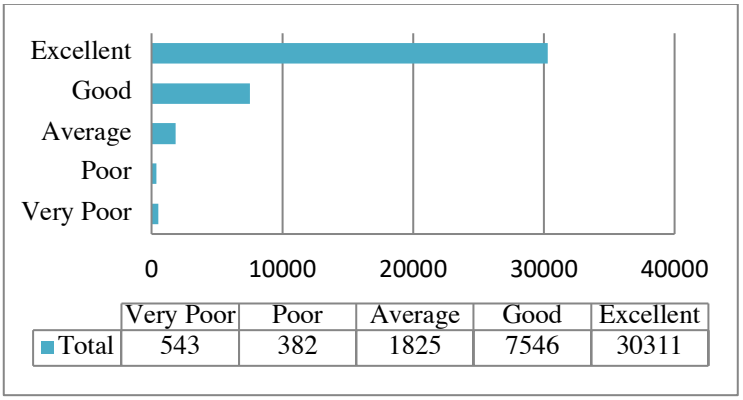

Figure 2. The reviews submitted based on rating

Figure 2 shows that the highest amount of review received by sellers was Excellent, while the least received was poor category with only 382 reviews. During the process of labelling, however, there were many reviews description the description in the content and the rating that didn't go into the same direction. Therefore, data labelling based on the content of reviews is needed.

4.2 Data Pre-Processing

Textual data pre-processing consists of spelling normalization, case folding, tokenizing, filtering, and stemming. Spelling normalization is the process of transforming the misspelled or non-standard words into standard words. After transformed into standard words, textual data are needed to be converted into one form. Case Folding is the process of converting characters into lowercase and removing all characters except a-z in text document. The next phase is tokenizing which split the sentence into pieces of words called tokens. Tokens are separated by using the white space as delimiter. After being splitted, stopwords algorithm will be implemented in order to eliminate irrelevant and less important

words, this phase is called filtering. The last phase is stemming that aims to convert words to their base form.

4.3 Data Labelling

After the pre-processing, the total number of reviews decreased to 23,639, which then will be automatically labelled by calculating the sentiment score using the Lexicon-Based sentiment analysis. The result of data labelling can be seen in the table below:

Table 2. Total reviews of each sentiment

Sentiment	Total Number of Reviews
Positive	19,298
Neutral	3,021
Negative	1,320

This analysis ignored the neutral class group, hence there were only 20,618 reviews left consisting of 19,298 positive reviews and 1,320 negative reviews. The formula of the sentiment score used in data labelling can be described with the following equation:

$$\text{SCORE} = \text{Total number of positive words} - \text{Total number of negative words} \quad (4)$$

4.4 TF-IDF Weighting

Term weighting is used in order to generate the classification model. The TF-IDF weighting can be seen in the table below.

Table 3. The Term Frequency-Inverse Document Frequency (TF-IDF) weighting

review	term							
	1	**2**	**3**	**4**	**5**	**6**	...	**5,305**
1st	0.2533	0	0.4271	0	1.0382	0.4998	...	0
2nd	0.2533	0.5354	0	0	0	0	...	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
20618th	0.7597	0	0	0	0	0	...	0

4.5 Visualization

4.5.1 Positive Review

In the positive reviews classification with 19,298 reviews, some of the most frequently used terms include: term 1 or "barang" with a total of 12,337, term 2 or "cepat" with a total of 8,531, term 3 or "sesuai" with a total of 6,596 , term 4 or "bagus" with a total of 6,489, and others which can be seen in figure 6. This collection of most frequently used words can be visualized using word cloud. Based on the word cloud of positive reviews classification, it can be concluded that the Tokopedia customers were satisfied with the fast response service, good quality product, and good packing quality.

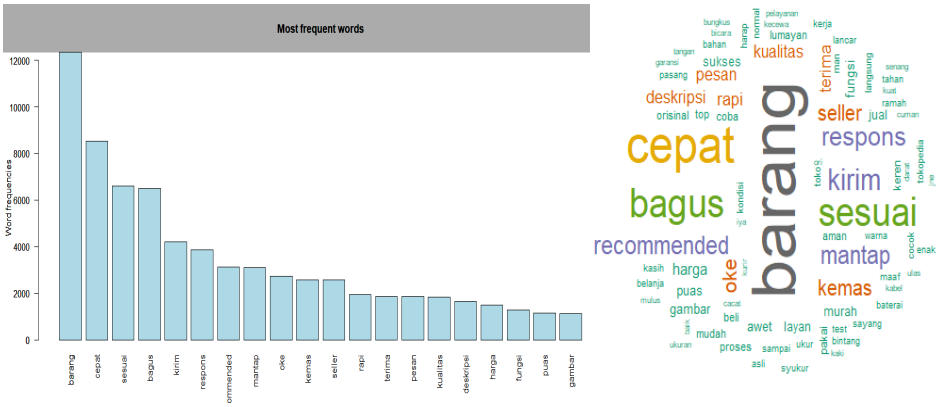

Figure 3. Visualization of positive reviews

4.5.2 Negative Review

In the negative classification with 1320 reviews, some of the most frequently used terms include: term 1 or "barang" with a total of 653, term 2 or "warna" with a total of 315, term 3 or "kirim" with a total of 309, term 4 or "gambar" with a total of 208, and others which can be seen in figure 7. This collection of most frequent words can be visualized using word cloud. Based on the word cloud of negative reviews classification, it can be concluded that Tokopedia customers were disappointed by the product delivery, product quality, and the mismatch between the description and the picture.

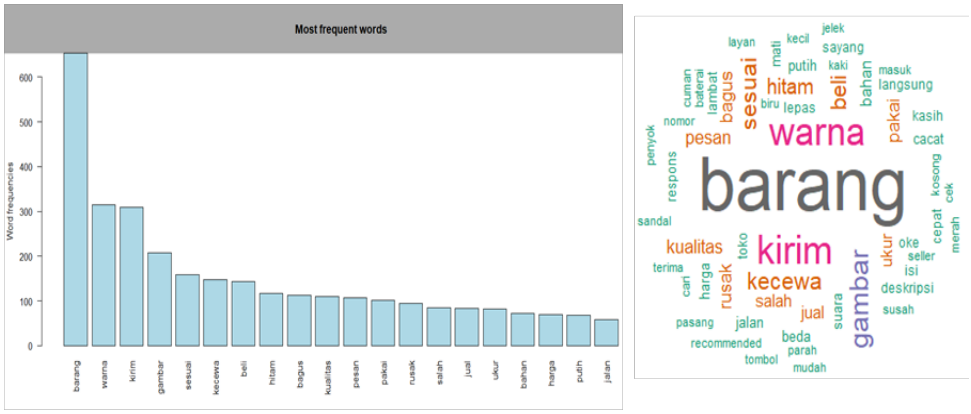

Figure 4. Visualization of negative reviews

4.6 Classification

4.6.1 Training and Test Sets

Training datasets and test datasets used in this analysis are labelled datasets, with the ratio of training data to testing data at 70%:30%.

Table 4. Proportion of training and test sets

Classification	Positive	Negative	Total
Training Data	13502	930	14432
Testing Data	5796	390	6186
Total	19298	1320	20618

4.6.2 Random Forest Classification

Based on the calculation using the formula of $m=\left(\frac{1}{2}|\sqrt{p}|,\ |\sqrt{p}|, 2|\sqrt{p}|\right)$ with a total of 5,305 variables, the mtry obtained is 36, 73, and 146 which will be used to generate a plot in order to observe how the mtry change affects the accuracy. The next step is to find the optimal value for ntree. In this analysis, there are 6 values of ntree that will be tuned: 10, 50, 100, 150, 200, dan 300.

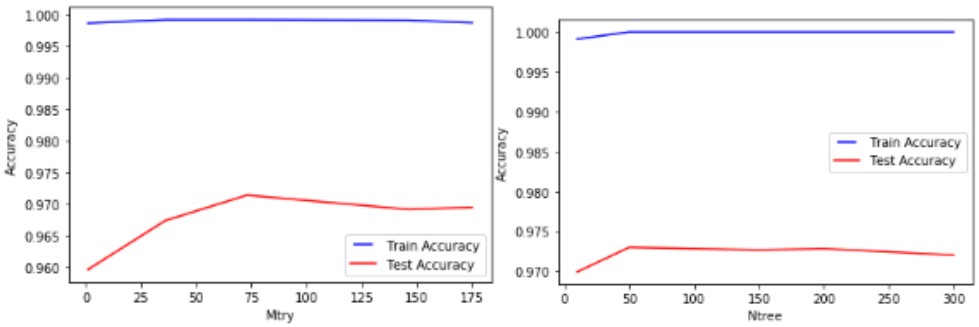

Figure 5. Plot of mtry and ntree

Figure 8 shows that the optimal mtry and ntree are 73 and 50 respectively because it has the highest accuracy compared to the others. Those parameters will be used to generate random forest model in order to predict the training sets and find the accuracy for testing sets or new data.

Table 5. Confusion matrix for the classification of testing sets

Actual	Predicted	
	Positive	Negative
Positive	5778	18
Negative	144	246

Table 9 shows that out of 5,790 positive review predictions, 5,778 reviews were grouped correctly into positive classification and 18 reviews were improperly grouped as negative. While for the negative reviews predictions, 246 reviews were correctly grouped and 144 reviews were incorrectly grouped, out of the 390 reviews. Therefore, the accuracy can be obtained from the confusion matrix using the following formula:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{(\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN})} = \frac{5778 + 246}{(5778 + 18 + 246 + 144)} = 0.9738$$

The accuracy of testing data is 0.9738 or 97.38% which means that out of 6,186 predicted reviews, there were 6,024 reviews correctly classified.

In order to evaluate the model, this analysis is using 10-Fold Cross Validation that observe 10 machine learning with different training and data sets to find predictions with the highest accuracy.

Table 6. Comparison of each machine learning accuracy

Machine Learning	Accuracy
1	97.09%
2	97.23%
3	96.47%
4	97.16%
5	96.88%
6	97.51%
7	96.88%
8	96.88%
9	96.67%
10	97.71%
Average	97.05%

Table 10 shows that out of the 10 observations on Random Forest, the 10th machine learning has the highest accuracy of 97.71%. In order to measure the performance of the classification model, this analysis was using the 10-Fold Cross Validation with the average accuracy obtained at 97.05%.

5 CONCLUSION

Tokopedia customers have good expectations for their products and the majority of the customers has purchased in electronics category. The accuracy of sentiment classification using Random Forest method with 70% training data and 30% testing data is 97.38%. The parameter used in this analysis is mtry = 73 dan ntree = 50. Generally, the visualization of positive reviews and negative reviews shows that most users give positive assessment such as fast response service, good quality product, and the packing are nice. While for the negative assessment, Tokopedia customers bring up their disappointment on delivery, product quality, and the mismatch between the description and the picture.

References

1. CupoNation, *10 Situs Belanja Online Terpopuler di Indonesia*, accessed 13 Maret 2020, https://www.cuponation.co.id/magazin/situs-belanja-online-terpopuler-sepanjang-tahun-2019 (2020)
2. B. Sartono, U.D Syafitri, *Metode Pohon Gabungan: Solusi Pilihan untuk Mengatasi Kelemahan Pohon Regresi dan Klasifikasi Tunggal*, Forum Statistika dan Komputasi **15**, 1-7 (2010)
3. E.M. Sipayung, H. Maharani, I. Zefanya, *Perancangan Sistem Analisis Sentimen Komentar Pelanggan Menggunakan Metode Naive Bayes Classifier*, Jurnal Sistem Informasi (JSI) **8**, 958-965 (2016)
4. R. Feldman, J. Sanger, *The Text Mining Handbook. New York: Cambridge University Press* (2007)

5. E.B. Santoso, A. Nugroho, *Analisis Sentimen Calon Presiden Indonesia 2019 Berdasarkan Komentar Publik di Facebook*, Jurnal Eksplora Informatika **9**, 60-69 (2019)
6. T. Tokunaga, M. Iwayama, *Text Categorization based on Weighted Inverse Document Frequency* (1994)
7. L. Breiman, *Random Forests*. Machine Learning **45**, 5-32 (2001)
8. L. Breiman, A. Cutler, *Manual on Setting Up, Using, and Understanding Random Forest 4.0* (2003)
9. P. Pitria, *Analisis Sentimen Pengguna Twitter Pada Akun Resmi Samsung Indonesia Dengan Menggunakan Naïve Bayes*, Jurnal Ilmiah Komputer dan Informatika (KOMPUTA) Universitas Komputer Indonesia (2014)
10. A. Deolika, Kusrini, E.T. Luthfi, *Analisis Pembobotan Kata pada Klasifikasi Text Mining,* Jurnal Teknologi Informasi **3**, 179-184 (2019)