Comparison Performance of Naive Bayes Classifier and Support Vector Machine Algorithm for Twitter's Classification of Tokopedia Services

R Kusumawati¹, A D 'arofah² and P A Pramana³

¹, ², ³Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia

E-mail: rosita_kusumawati@uny.ac.id

Abstract. Tokopedia is one of the online shopping centers in Indonesian that carries the business model marketplace. Positive and negative opinions in Twitter from Tokopedia users about company services are source of information for the management. Naive Bayes Classification (NBC) and Support Vectore Machine (SVM) are techniques in data mining used to classify data or users opinion. The algorithm of NBC is very simple since it only use text frequency to compute the posterior probability for each classes. While SVM algorithm is more complex than NBC. SVM develop hyperplane equation which separate data into classes perfectly. The researcher wants to compare the performance of the NBC and SVM algorithms and use them to classify user opinions on Tokopedia’s services, because these two algorithms have different approaches and difficulty levels. Classification included positive and negative class only. Accuracy, precision and recall value are used to compare the performance of both algorithms. Research evaluation shows that SVM linear kernel technique outperform NBC technique with the accuracy 83.34%.

1. Introduction
Tokopedia is one of the e-commerce startups with large assets that accommodate sellers and buyers to make transactions quickly and easily. Whereas Twitter is a microblog based social network with 24.34 million users in May of 2016 that allows users to send and read text based messages known as “tweets” [11]. Twitter can be a source of text data of opinion and community sentiment on Tokopedia’s services which can be analysed for the purposes of an organization or a company. A sentimental analysis, also called an opinion mining, is a field of study that that not only analyse people's opinions, but also people's sentiments, evaluations, judgments, attitudes, and emotions on entities such as products, services, organizations, individuals, problems, events, topics, and other attributes [5]. These unstructured opinions text data can be classified using various text processing methods called text mining [12].

Algorithms used to classify opinions include Naive Bayes Classifier (NBC) and Support Vector Machine (SVM). The NBC is a simple probabilistic-based method which classify opinion based on the maximum value of the posterior probability obtained through the probability of each class (prior probabilities) and the probability of each word occurring (conditional probability) in the training data. Although very simple, NBC has a high degree of accuracy and performance in classifying data text [10].
The Support Vector Machine (SVM) algorithm is one of machine learning (supervised learning) method which can classify opinion by searching for the best dividing field or hyperplane that separates high-dimensional text data perfectly into classes [14]. Hyperplane can be found by maximizing the margin or the distance between the closest class points (support vector) and the hyperplane. However, samples data are often not linearly separated, the SVM introduces the idea of increasing data dimensions. Typically, the use of higher space dimensions will causes engine problems and overfitting. The problem can be solved by the use of dot-product in space [3].

Although many research on text mining have been done, but there are still many issues about the performance of the algorithm that need to be addressed further. This article will discuss the application of Naive Bayes Classifier and Support Vector Machine algorithm to classify tweets of Tokopedia’s Services and compare the performance of the two algorithms. The remainder of this paper is organized as follows, section 2 gives works related to text mining using NBC and SVM. The concepts of NBC and SVM will be explained deeper in section 3. We will use the accuracy, precision and recall level which will explained further in section 4 to compare the performance of each algorithm. The classification steps of Tokopedia’s services starts from collection of data, labelling preprocessing, and sharing then NBC and SVM classification process will be discussed in section 5.

2. Related Work
Numbers of research on sentiment analysis using NBC and SVM algorithm have been done. Narayanan et.al conducted a sentiment analysis study to find out the film review opinions by Indian audiences using NBC classification [7]. The film reviews in this study come from the Internet Movie Database (IMDb). The results of the research quickly and accurately with the level of accuracy 88.80% of 25,000 opinions. While Wahyuningtyas use NBC algorithm also to classify spam and not spam tweets [13]. The classification accuracy of spam and non-spam tweets is 95.57%. Based on the results of the research can be known words that often appear on the class spam is bahasa, follow, and inggris. Research to analyse SVM performance has been done to classify English opinion about self-driving cars and apple products using tweet data divided into six classes using WEKA program [1]. The accuracy, precision and recall values of each topic are 59.91%, 70.8%, 84.1% for self-driving cars problem and 71.2%, 70.2%, 71.2% for apple products analysis. Pratama et.al also used SVM algorithm for text mining of Speedy Telkomsel subscribers’ complaint in twitter using feature selection combinations i.e. term frequency, document frequency, information gain and chi-square) and Gaussian RBF kernel function. And the resulted accuracy value for term frequency feature is of 82.50% [8].

3. Problem Formulations and Methodology

3.1. Naïve Bayes Classifier (NBC)
The Naïve Bayes Classifier is a supervised learning technique in a form of probabilistic classification based on Bayes theorem with intermediate naive assumption. Where the Bayes rules can be stated as follows:

\[
P(B_j|A) = \frac{P(B_j)P(A|B_j)}{P(A)}
\]  

(1)

Using the Bayes rule in equation (1), \(P(c_i|x_1, ..., x_n)\) the conditional probability of class \(c_i\) given the word feature \((x_1, ..., x_n)\) in a particular can be expressed as follows:

\[
P(c_i|x_1, ..., x_n) = \frac{P(x_1, ..., x_n|c_i)P(c_i)}{P(x_1, ..., x_n)}
\]  

(2)

where \(P(x_1, ..., x_n|c_i)\) is the conditional probability of the word feature \((x_1, ..., x_n)\) in a particular class \(c_i\) and \(x_i\) state the number of words \(i\) appeared in a text. In the NBC algorithm for text mining problem, the assumption guarantee that the appearance of a word in a text does not affect the
appearance of other words [2]. In the other words, the NBC algorithm assumes that variables $x_1, ..., x_n$ are mutually independent.

The Naïve Bayes Classifier algorithm can be divided into two types i.e. multivariate Bernoulli and multinomial Naïve Bayes [6]. This study used the model of multinomial Naïve Bayes since it assumed the mutual independence of each word for all classes and $P(x_1, ..., x_n) = 1$ or constant, so equation (2) can be written as follows:

$$ P(c_i|x_1, ..., x_n) = P(x_1, ..., x_n|c_i)P(c_i). \quad (3) $$

The posterior probabilities values can be obtained based on the probability of each class (prior probabilities) and the probability of each word (conditional probabilities) in the training data by simplifying equation (3) as follows [9].

$$ P(c_i|x_1, ..., x_n) = \prod_{j=1}^{n} P(x_j|c_i) P(c_i) \quad (4) $$

where

- $i$: positive class, negative class
- $n$: number of word features in the training data
- $P(c_i|x_1, ..., x_n)$: probability of words $x_1, ..., x_n$ in class $c_i$ (posterior probabilities)
- $P(x_j|c_i)$: probability of occurrence of word $x_j$ in class $c_i$ (conditional probabilities)
- $P(c_i)$: probability of class $c_i$ in the training data (prior probabilities)
- $(x_1, ..., x_n)$: word features

The equation (4) is a probability model of the Naïve Bayes theorem which is used in the classification process. In the Naïve Bayes Classifier, the testing data enter the class $c_i$ that has a maximum posteriori (MAP) or $c_{MAP}$. The calculation of $c_{MAP}$ value is defined as follows:

$$ c_{MAP} = \arg\max_{c_i \in C} \prod_{j=1}^{n} P(x_j|c_i) \quad (5) $$

with the prior probability values as follows:

$$ P(c_i) = \frac{N_{c_i}}{N} \quad (6) $$

where $N_{c_i}$ is the amount of training data that has class $c_i$ and $N$ is the number of data used in the training data.

The conditional probabilities values as follows:

$$ P(x_j|c_i) = \frac{n_j}{n} \quad (7) $$

where $n_j$ is the number of occurrences of the word $x_j$ in class $c_i$ while $n$ is number of words contained in class $c_i$.

Sometimes there are words that never appear in any of the classes during the classification process so that the resulting $P(x_j|c_i)$ value is zero. To prevent the occurrence of division by zero, then Laplace smoothing is used by adding the word frequency as much as 1(add-one) so that the calculation of $P(x_j|c_i)$ becomes

$$ P(x_j|c_i) = \frac{n_j + 1}{n + n_k} \quad (8) $$

where $n_k$ are the number of different (unique) words that appear in the training data.

### 3.2. Support Vector Machine (SVM)
A text or tweet can be expressed as set $x_i$, where $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$. Suppose a given set $x = \{x_{i1}, x_{i2}, ..., x_{in}\}$ with $x_i \in \mathbb{R}^p$ has a certain pattern, that can be grouped into positive class and negative class. Thus, each datum and class label can be denoted as $y_i \in \{-1, +1\}$ so the data are pair $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, then it is assumed that the data are perfectly separated (linear separable) by the $p$ dimensional separator function called hyperplane $H_0$ where $n$ is the number of data that can be defined in the equation (9) as follows,

$$w \cdot x_i + b = 0$$

(9)

where $w$ is weights vector and $b$ is scalar. The illustration of hyperplane and support vector can be seen in Figure 1.

![Hyperplane and support vector](image)

**Figure 1.** Hyperplane and support vector

The problem of hyperplane $H_0$ is the same as looking for the best split field with the largest margin value which can be formulated into equation (10) as follows,

$$\min \frac{1}{2} ||w||^2$$

(10)

with the constraints given by,

$$y_i (w^T \cdot x_i + b) \geq 1, \ i = 1, ..., n$$

The classification of the above linear case data explains that data can be separated into two classes perfectly. For non-linearly separable data, the SVM formula should be modified because the equation constraint cannot be satisfied and the optimization cannot be performed. Therefore, slack variable $s_i$ ($s_i \geq 0, \forall i; s_i = 0$ if $x_i$ correctly classified) needs to be added so that $y_i (w^T \cdot x_i + b) \geq 1 - s_i$ is obtained. The problem in equation (10) can be formulated into functions (11) as follows,

$$\min \frac{1}{2} ||w||^2 + C (\sum_{i=1}^{n} s_i)$$

(11)

with constrains,

$$y_i (w^T \cdot x_i + b) \geq 1 - s_i , \ i = 1, ..., n$$

(12)

and $s_i \geq 0$ for $\forall i$.

(13)

$C$ plays a role in minimizing training errors and reduce complexity in the model. Using the Lagrange function for optimization problem, the optimization problem in equations (11) – (13) can be stated as optimization problem without constraints in equation (14) as follows,

$$\min L_p (w, b, s, \alpha) = \frac{1}{2} ||w||^2 + C (\sum_{i=1}^{n} s_i)$$

$$+ \sum_{i=1}^{n} \alpha_i (1 - y_i (w^T \cdot x_i + b) - s_i)).$$

(14)

Non-negative $\alpha_i$ variables is called Lagrange Multiplier where $\alpha_i \geq 0$. The objective function of equation (14) is to minimize $L_p$ to $w$ and $b$, and the same time maximize $L_p$ to $\alpha$. By utilizing partial derivatives $L_p$ to $w, b$ and $s$, the dual problem of equation (14) is as follows,
maks \[ L_d = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i, x_j) \]  \hspace{1cm} (15)

with constrains,

\[ 0 \leq \alpha_i \leq C , \text{with } i = 1, \ldots, n \]  \hspace{1cm} (16)

and \[ \sum_{i=1}^{n} \alpha_i y_i = 0 \]  \hspace{1cm} (17)

4. The Research Method

Research on sentiment classification evaluation is performed to test the results of the classification by measuring the performance value of the system that has been made. Test parameters used for evaluation are accuracy, precision and recall whose calculations are obtained from the confusion matrix table. The calculation of accuracy, precision, and recall is obtained through the formula [10] :

\[ \text{Akurasi} = \frac{TP+TN}{TP+FP+TN+FN} \times 100\% \]  \hspace{1cm} (18)

\[ \text{Presisi positif} = \frac{TP}{TP+FP} \times 100\% \]  \hspace{1cm} (19)

\[ \text{Recall positif} = \frac{TP}{TP+FN} \times 100\% \]  \hspace{1cm} (20)

\[ \text{Presisi negatif} = \frac{TN}{TN+FN} \times 100\% \]  \hspace{1cm} (21)

\[ \text{Recall negatif} = \frac{TN}{TN+FP} \times 100\% \]  \hspace{1cm} (22)

Data used in this study are Twitter data which are obtained using the Twitter API (Application Programming Interface). There are 120 tweets used in the analysis. The first steps in the data collection is to take tweet data from the Twitter API by using the "twitteR" package on R and the searchTwitter function. All stages of the research can be seen in Figure 2.

Figure 2. The stages of research

5. Research, Analysis, and Discussions

5.1. Data Collection

Stage of data collection is retrieving or crawling tweets data from the Twitter API (Application Programming Interface) by using the "twitteR" package on the R program and using the searchTwitter function. There are many attributes in the retrieval Twitter data, but this study use only the text attributes. Twitter data collection steps using the Twitter API are the following:

5.1.1. Connect R Program with Twitter API

Making application management on Twitter is required to connected R programs with the Twitter API. In the application management application obtained consumer key (API key), consumer secret (API
secret), oauth token, and oauth secret token required to send a secure authorized requests to Twitter API for R program can be used to extract Twitter data.

5.1.2. Data Crawling in Twitter
Data crawling in Twitter is as much as 300 data tweets with the word Tokopedia as a keyword. Tweet data retrieval using the searchTwitter() function with parameters (keyword, n). The keyword parameter is filled with the word "@tokopedia", and the sample size n is filled with the desired number of tweets that is n = 300. After the data retrieval process, then the tweet data obtained is converted into the data frame and stored in file.csv format.

5.2. Data Labeling
After crawling the data and then data filtering. Data filtering aims to collect data containing only the user's opinion of Tokopedia's service. Data tweets containing ads and considered irrelevant with Tokopedia removed and leave 120 data tweets from 300 tweets. Furthermore, data labeling are done by giving positive and negative labels on each tweet. The negative tweets and positive tweets are 53% and 47% tweets respectively from all tweets.

5.3. Preprocess
Tweet data that have been labeled then done pre-process stage. The preprocessing stage consists of five stages:

5.3.1. Case Folding
Case folding is the step of converting all the characters into lowercase in the document.

5.3.2. Tokenizing
Tokenizing is the breaking step of a text document into multiple tokens or words. In the tokenizing stage is also done the removal of punctuation, number, mention, hashtag, and url. The purpose of tokenizing stage is to be able to calculate the frequency of each word that appears.

5.3.3. Normalization
Normalization is the phase of word improvement process by converting the word abbreviation into standard word form. The list of word abbreviations is stored in the form of a database using MySql. In this study the program code obtained from previous research conducted by Khotimah (2014).

5.3.4. Stopwords Removing
Stopwords is a list of words that are considered unimportant or irrelevant to content on tweet. Removal of stopwords for indonesian words using stopwords database (Tala, 2003) amounting to 759 words and other additional words that are not meaningful on tweet.

5.3.5. Stemming
The process of stemming is the process of converting words into basic and non-standard words into standard words. In this study, stemming was done using the algorithm Nazief and Adriani (1996). The program code used in the study is the program code followed by the Khotimah study (2014). Stemming stage is the last stage in the pre-process stage.

After the pre-processing phase, the token result will be the term. Term is a unique token (word). The next stage is the creation of term document matrix that aims to determine the frequency of occurrence of words in the document. The weighting of words performed in this study is to use the term frequency (tf) by looking at the number of terms that appear in each tweet.

5.4. Data Sharing
After pre-processing, the tweet data that have gone through pre-processing stage are then divided into two parts, namely training data and testing data. In this study the training data and testing data used are formulated using 80:20 rule. The amount of training data and testing data are 96 and 24, respectively.

After data training and data testing have shared, then that is making wordcloud representation by using term document matrix which have been done in document. Words that appear on positive tweets and negative tweets are represented by wordcloud. The results of the positive class can be seen in Figure 3 and the word in the negative class can be seen in Figure 4.

![Wordcloud representation in positive class](image)

![Wordcloud representation in negative class](image)

In Figure 3 it can be seen that the word that often appears in the positive class is bagus, sale, ramadhan, diskon and in Figure 4 it can be seen that the words that often appear in the negative class is tidak, bayar, transaksi, kirim. from the processing using the pre-process generated as many as 388 terms that will be used as many features to perform classification on both algorithms.

5.5. Classification

5.5.1. Classification using naive bayes algorithm.

The method of classification in this research is by using Naïve Bayes algorithm. Based on the results of the pre-processing, it will then calculate the probability of occurrence of each class on the sample (prior probabilities) and the per-word weighting chances to be classified into positive classes and negative classes (conditional probabilities). The number of terms generated from the pre-process will be used on the NBC algorithms. There are as many as 388 terms generated. The classification model is the process of classifying with test data using training knowledge derived from the training data to classify the test data. The probability of occurrence for each class (prior probabilities) in the training data is as follows:

\[ P(c_{positive}) = 0.4166667 \] and \[ P(c_{negative}) = 0.5833333 \]

The results of word probability calculations in each class (conditional probabilities) using Naïve Bayes Classifier algorithm can be seen in Table 1.
Table 1. Word probability of each class (conditional probabilities)

| Words         | Negative     | Positive    |
|---------------|--------------|-------------|
| Terimakasih   | 0.01785714   | 0.1250000   |
| Tidak         | 0.39285710   | 0.1250000   |
| Tokopedia     | 0.07142857   | 0.4500000   |
| Proses        | 0.05357143   | 0.0250000   |
| Bagus         | 0.00000000   | 0.4250000   |
| Transaksi     | 0.08928571   | 0.0000000   |
| Tolong        | 0.10714290   | 0.0000000   |
| Iklan         | 0.00000000   | 0.2500000   |
| Ramadhan      | 0.00000000   | 0.1250000   |
| Flash         | 0.1785714    | 0.0000000   |
| Sale          | 0.1964286    | 0.1250000   |

For example there is a tweet "Iklan bagus Tokopedia" with a posterior probability value on each is

\[ P(\text{pos}|\text{tweet}) = P(\text{iklan}|\text{pos})P(\text{bagus}|\text{pos})P(\text{tokopedia}|\text{pos})P(\text{pos}) \]
\[ = (0.250)(0.425)(0.450)(0.4167) \]
\[ = 0.0199 \]

and

\[ P(\text{neg}|\text{tweet}) = P(\text{iklan}|\text{neg})P(\text{bagus}|\text{neg})P(\text{tokopedia}|\text{neg})P(\text{neg}) \]
\[ = (0.000)(0.000)(0.071)(0.583) \]
\[ = 0.000 \]

Because \( P(\text{positive}|\text{tweet}) > P(\text{negative}|\text{tweet}) \) then tweet it into the positive class. In the next stage of testing the results of classification using test data that has been determined. The results of the prediction of the classification with Naïve Bayes algorithm can be used to determine the level of accuracy of a model.

5.5.2. Classification using support vector machine algorithm. Classification modeling is done using Support Vector Machine algorithm. The kernel function used linear kernel using one parameter \( C(\text{cost}) \). \( C(\text{cost}) \) is the parameter of the penalty of the error in the classification and the value is determined by the researchers. In this study the user determines the value \( C \), the values are 0.1; 1; and 10 for modeling the data classification of training and later searching for the best \( C \) that can increase the value of accuracy. Data will be classified by building SVM model using linear kernel function through \textit{packages RtextTools} and \textit{e1071} with \textit{create_container} and \textit{train_model} as modeling function with the training data. By using the function \textit{train_model} can know the number of support vector contained in the data train, and in this research obtained as many as 81 support vector. The determination of the previous vector support has been done by the system to determine the function, the form of the value of \( \bar{w} \) and \( b \). The values of \( \bar{w} \) and \( b \) will be used to find identical hyperplane. After obtaining the value with the calculation, the determination of the classification result using the SVM model already constructed using this training data will be used to process the data. The number of terms generated from the preprocess will be used on the SVM algorithm, which is similarly used on previous NBC algorithms. There are as many as 388 terms generated. The classification model is the process of classifying with test data using training knowledge derived from the training data to classify the test data. Establishment of the classification model in the test data using the function classify\_model. In the evaluation and validation stage of the results, the Confusion Matrix table is used to evaluate the classification results (predictions) in the test data. From the calculation of accuracy
value got the best accuracy on SVM is with value of C, that is 0.1. The example of SVM for this research can be seen in Table 2.

**Table 2. Examples of false tweets predicted in SVM**

| Tweets                                      | Actual | Predicted |
|---------------------------------------------|--------|-----------|
| wakakak wajar temen ku depet flash sale untung na | positive | negative |
| maklum ada yg tidak dapet produk batas akses puluh | positive | negative |

5.5.3. **Evaluation**. The calculation of accuracy, precision and recall value for each data type for SVM and NBC algorithm can be seen in Table 3 below.

**Table 3. Accuracy, precision and recall value with NBC and SVM**

| Algorithm | Accuracy | Precision | Recall |
|-----------|----------|-----------|--------|
|           |          | positive  | negative | positive  | negative |
| NBC       | 75%      | 100%      | 33.33%  | 62.50%    | 100%     |
| SVM       | 83.34%   | 100%      | 75%     | 66.67%    | 100%     |

Table 3 provides the evidences that the SVM does better job that the NBC for classifying opinion on Tokopedia’s services. All criteria show higher values for the SVM and the NBC.

6. **Conclusion**

According to the accuracy, the precision and the recall value, the performance of SVM algorithm in classifying Tokopedia’s services is much better than NBC. The worldcloud suggests positive result in Tokopedia’s quality product and discount or sale program. The negative class corresponds to the shipment and the payment procedures.

**References**

[1] Ahmad, M., Aftab, S., & Ali, I. 2017. Sentiment analysis of tweets using SVM. *International Journal of Computer Applications*, 25-29.

[2] Apriliyanti, A. 2015. Sentiment analysis with naive bayes to see people’s perception of batik on twitter social network. *Proceedings of the National Seminar Mathematics and Mathematics Education of Muhammmadiyah Surakarta University*, 836.

[3] Bowell, D. 2002, August 6. *Introduction to support vector machines*. taken from dustwell.com: dustwell.com/PastWork/IntroToSVM.pdf

[4] Khotimah, H. 2014. Pemodelan hybrid tourism recommendation menggunakan hidden markov model dan text mining berbasis data sosial media. *Thesis of Institute Pertanian Bogor*

[5] Liu, B. 2012. *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers.

[6] Manning, C.D., Raghavan, P., & Schutze, H. 2009. *An introduction to information retrieval*. Cambridge University Press: New York.

[7] Narayanan, Vivek, Arora, I., & Bhatia, A. 2013. *Fast and accurate sentiment classification using an enhanced naive bayes model*. Varanasi, India: Indian Institute of Technology.

[8] Pratama, E., & Triliksono, B. 2015. The classification of customer complaint topics based on tweets using the combination of extracted features in the method of support vector machine (SVM), *Jurnal Edukasi dan Penelitian Informatika (JEPIN)* 1(2), 53-59.

[9] Raschka, S. 2014. *Naive bayes and text classification i - introduction and theory*. Birmingham: Packt Publishing.

[10] Routray, P., Swain, C. K., & Mishra, S. P. 2013. A survey on sentiment analysis. *International Journal Of Computer Applications*, 2-4.

[11] Statista. July 2016. *Number of active twitter users in leading markets as of may 2016 (in millions)*. Taken from https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/ on 25 March 2018.
[12] Vijayarani, S., Ilamathi, J., & Nithya. 2015. Preprocessing techniques for text mining - an overview. International Journal of Computer Science & Communication Networks, 5(1), 7-16.

[13] Wahyuningtyas, A. 2016. Spam detection on twitter using naïve bayes algorithm. Bogor Agricultural University.

[14] Zainuddin, N., & Selamat, A. 2014. Sentiment analysis using support vector machine. IEEE 2014 International Conference on Computer, Communication, and Control Technology, 333-337.