Comparison of Feature Weighting in SVM Performance for Sentiment Analysis of Jakarta BRT

Widyawan, Nourma Reizky Damayanti, Teguh Bharata Adji, and Guntur Dharma Putra
Department of Electrical Engineering and Information Technology, Faculty of Engineering, Universitas Gadjah Mada, Yogyakarta, Indonesia

Abstract. Twitter has been utilized to distribute opinions from netizens directly to public service providers, such as Jakarta Bus Rapid Transit (BRT), in an efficient and effective way. In this context, the opinions formatted in textual data can be analyzed to help BRT operators improve their facilities and services via sentiment analysis, which consists of multiple steps: preprocessing, feature weighting, classification, and evaluation. The preprocessing and feature weighting are the key processes that may significantly affect the classification algorithm performance. Several researches have investigated these key processes, specifically to observe its effect in classification performance. However, none of those researches compare n-gram feature tokenization with feature weighting in Bahasa Indonesia. The present study compares the combination of n-gram feature tokenization with feature weighting to the performance of Support Vector Machine algorithm. The present study utilizes TF-IDF, TF-CHI, TF-RF, and TF-OR as the feature weighting scheme. The results show that TF-IDF has the highest performance of 79.3% (accuracy), 83.2% (precision), and 83.6% (recall), and 82.2% (F1 score).

1. Introduction
Bus Rapid Transit (BRT) is a form of mass rapid transportation implemented in several developed cities, such as Jakarta since 2004. Additionally, BRT can also reduce traffic congestion as passengers were previously using private vehicles for their daily transportation. According to the first semester report of 2016, there were 8.51, 8.15, 9.01, 9.10, 10, and 10.206 millions of passengers in January, February, March, April, May, and June, respectively [1]. The value was even increased to 11 millions of passengers in August the same year [2]. The increase of passengers must also be balanced with facilities and services upgrades.

Jakarta BRT operator has distributed questionnaires to obtain insights regarding passenger satisfaction and opinion about BRT facilities and services. However, the effort requires significant cost, while there is also another effort to obtain insights via social media, a means to distribute people’s opinion in many cases. Social media users of Indonesia is categorized in top rank, which is nearly 40% of total social media users [3] with Twitter as the most popular social media. In 2015, Twitter user in Indonesia dominates the twitter user, ranked number 1 in country list. Twitter users consist of private, organization, community, and public service accounts, including BRT operator in Jakarta. The official account of BRT operator can be utilized for bi-directional communication to BRT users, as they may submit comments or complaints directly.

BRT users' opinions submitted via Twitter are referred as tweets, in the form of textual data, which can be analyzed to obtain information for improving services or other text mining purposes. Sentiment
analysis, or opinion mining, is a branch of text mining that aims to classify texts into groups of negative or positive sentiments. Generally speaking, there are two types of sentiment analysis, namely supervised and unsupervised analysis, or learning. Supervised learning is a sentiment analysis which uses predefined training data to classify sentiments of texts, while unsupervised learning is the one without the need of training data. Supervised learning tends to have higher accuracy as it has a knowledge base (training data) which the classification model relies on [4].

Before creating the classification model for sentiment analysis, the textual data must go through pre-processing stage. One of which is feature tokenization that groups texts into several tokens. N-gram is an example of feature tokenization, which is known to have an effect on text classification performance [5]. In sentiment analysis, n-gram helps to recognize the negative or positive sentiments for the phrase.

In addition to n-gram, the classification performance may also be affected by feature weighting, a stage which assigns weight to extracted features with respect to their significance. Feature weighting is important as it provides the basis of information extraction, such as the feature appearance frequency and distribution, and the number of documents in each class [6]. Taufikurrhaman compared various types of feature weighting and investigated its effect on classification performance [6]. However, his research only employed English as the only language and unigram as the only feature tokenization.

Several works have focused on sentiment analysis of Indonesian textual data. Some of those focus on accuracy comparison of different classification algorithms, such as one conducted by Yusuf [7], which compares the accuracy of the Naive Bayes and Support Vector Machine (SVM). Research conducted by Langi [8] and Rahmawati [4] compares various types of classification methods and analyze their performance, showing that SVM has better accuracy and performance. This is also the reason why SVM algorithm is chosen as the main classification method in our study.

To the best of our knowledge, there are no studies that compare the effect of feature weighting and n-gram tokenization on the performance of SVM for Indonesian textual data. Therefore, the present study aims to study the performance of SVM with various types of n-gram feature tokenization and several feature weighting schemes in Indonesian textual data, especially for sentiment analysis.

2. Theoretical Background

2.1. Twitter
Created by Jack Dorsey in 2006, Twitter is one of social media which provides services to its users for posting 140-character length of text messages, named tweet, to its follower. Twitter provides an Application Programming Interface (API) to access and retrieve tweets on Twitter accordingly.

2.2. Term Frequency Inverse Different Frequency Weighting Method
Term Frequency-Inversed Document Frequency (TF-IDF) is a method to calculate the weight of each feature in a document. If a feature frequently appears on a particular document, the weight of the feature gets bigger and the feature is considered important [9]. Term Frequency (TF) is marked by the frequency of occurrence of features in each document in the dataset. Inverse Document Frequency is a value to measure the occurrence of words in all documents in the dataset, as shown in Equation (1) [10].

\[ idf = \log \left( \frac{N}{df_t} \right) \]  

(1)

where \( df_t \) is the number of documents where \( t \) feature appears on and \( N \) is the total documents. The value of TF-IDF(\( W_t \)) can be calculated by multiplying \( tf_t \) and \( idf \) in equation (1), as formulated in equation (2) [10]:

\[ W_t = tf_t \times \log \left( \frac{N}{a+c} \right) \]  

(2)
2.3. Term Frequency Odds Ration Weighting Method

Term Frequency-Odds Ration (TF-OR) is a development of supervised learning methods in text categorization that requires data labelling prior to data categorization. TF-OR is the multiplication of OR value in equation (3) which then multiplied by the value of $tf_t$ [10].

$$OR_t = \log\left(\frac{a \times d}{b \times c}\right)$$

where $a$ is the number of documents in positive class where $t$ feature appears on, $b$ is the number of documents in positive class with no appearance of $t$ feature, $c$ is the number of negative class documents where $t$ feature appears on, and $d$ is the number of document of negative class with no appearance of $t$ feature.

2.4. Term Frequency Relevance Frequency Weighting Method

Term Frequency-Relevance Frequency (TF-RF), is a development of typical feature weighting methods, where the relevance frequency or $rf$ indicates that only certain documents that contains $t$ feature are used in this weighting scheme [10]. $rf$ can be formulated with equation (4) [10]. To get the TF-RF weight, $rf$ is multiplied by $tf_t$.

$$rf_t = \log\left(2 + \frac{a}{\max(1,c)}\right)$$

where $a$ is the number of positive class documents where $t$ feature appears on and $c$ is the number of negative class document where $t$ feature appears on.

2.5. Term Frequency Chi Square Weighting Method

Term Frequency Chi-Square is a weighting method in supervised learning. The calculation of Chi-Square value is shown in equation (5) [10]. To get the TF-RF weight, the value of equation (5) is multiplied by the value of $tf_t$.

$$X_t^2 = N \times \frac{(a \times d - b \times c)^2}{(a+c)(b+d)(a+b)(c+d)}$$

where $a$ is the number of documents in positive class where $t$ feature appears on, $b$ is the number of documents in positive class with no appearance of $t$ feature, $c$ is the number of negative class documents where $t$ feature appears on, $d$ is the number of document of negative class with no appearance of $t$ feature, and $N$ is the total documents.

2.6. Support Vector Machine Algorithm

Support Vector Machine (SVM) is a machine learning classification technique that was first introduced in 1992 by Boser. SVM overcomes overfitting problems and aims to divide the data linearly into two categories with linear hyperplane, shown in Figure 1. Support vectors or members of the closest training data category are simple examples for determining the hyperplane.
In Figure 1, \( d_1 \) and \( d_2 \) are hyperplane margins, which are measured to get the maximum point to determine the optimum hyperplane. A margin is the distance between the hyperplane and the closest data from both categories. SVM is chosen as a text classification method because it corresponds to the properties of the text [11].

3. Methodology

Figure 2 shows the research methodology flowchart of the present study.

![Research flowchart](image)

\textbf{Figure 2.} Research flowchart.

3.1. Twitter Data Collection

The dataset is a collection of tweets from Twitter extracted via Twitter streaming API, in which tweets that contain keywords about BRT is captured. The dataset is used as training and test data on the classification model.

3.2. Manual Labeling and Verification

The dataset, described in Section 3.1, is then manually labeled to two categories, namely "negative" and "positive". As for the ground truth, the labeled tweets are verified by Bahas Indonesia experts.
3.3. Preprocessing
The preprocessing stage aims to prepare the textual data before being classified and improve the quality of the text by eliminating unused elements or parts in the sentiment analysis process. Some preprocessing stages are as follows:

3.3.1. Case Folding. A process to uniform each character to small cases.
3.3.2. Cleansing. A step in which insignificant elements are removed. In the present study, web URLs, hashtags, username, or mentions are removed. Furthermore, some typos and abbreviations are treated accordingly.
3.3.3. Stemming. A process to stem the words to its root words, using Sastrawi stemmer [12].
3.3.4. Stopword Removal. A step to remove stopwords, common words that have no semantic information. The removal is based on Tala stopwords list [13].
3.3.5. Feature tokenization. A process for dividing documents into several groups of tokens, where each token contains one or more words. In the present study, several feature tokenizations are used, such as unigram, bigram, trigram, n-gram, or a combination of several feature tokenizations.

3.4. Feature Weighting
Feature weighting aims to give weight to features or tokens to determine the effect of tokens on documents or classes, represented in vector form. The vector representation is the result of the calculation of the feature weighting method. In the present study, the feature weightings are Term Frequency Inverse Document Frequency (TF-IDF), Term Frequency Relevance Frequency (TF-RF), Term Frequency Odd Ratio (TF-OR) and Term Frequency Chi-Square (TF-CHI). Each feature weighting method is combined with the n-gram feature tokenization.

3.5. Classification
The present study employs SVM algorithm. The SVM algorithm takes data, which are represented in vector from previous processes in the research methodology.

3.6. Evaluation
Evaluation is used to determine the performance of the classification result, done by using train-test split method that divides the dataset into training and test data. The evaluation results are represented in the confusion matrix table, in which four classification results are presented, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The performance is measured using four metrics, namely accuracy, precision, recall and F1-score.

4. Result and Discussion
4.1. Data Collection
The dataset consists of 623 tweets consisting of 339 positive and 284 negative tweets, manually labeled by Bahasa Indonesia experts. Positive tweets constitute good or satisfying comments about Jakarta BRT, while negative tweets constitute bad or unsatisfactory comments. An example of a tweet from the dataset is shown in Table 1.

| Sentiment | Tweets |
|-----------|--------|
| Negative  | @PT_TransJakarta lapor dong, d jalur koridor 2 arah kwitang ada jalur yg masi tertutup batas, tlg d buka tq |
| Positive  | @Ssteffy_ai93 Utk promo Transjakarta dgn Flazz BCA Ibu Steffy dpt klik link https://t.co/gD9TC2IG1b  Tks :) ^Olid |
4.2. Preprocessing

The tweets still need to be processed further to delete meaningless element and stopwords, and stem each word to its root form. An example of preprocessing is presented in Table 2.

**Table 2.** Preprocessing example results.

| Prior to preprocessing | After preprocessing |
|------------------------|---------------------|
| @PT_TransJakarta lapor dong, d jalur koridor 2 arah kwitang ada jalur yg masi tertutup batas, tlg d buka tq | lapor jalur koridor arah kwitang jalur tutup batas tolong buka tq |

As shown in Table 2, there are deletions of some elements and the conversion of abbreviations to the complete forms in Bahasa Indonesia. However, several words are still unaltered and not successfully converted into the correct phrases due to limitations in preprocessing dictionary. In addition, some stopwords are also deleted as it has no significant meaning.

4.3. Classification Evaluation Result

The preprocessing and feature tokenization are then followed by feature weighting and classification stage. There are nine types feature tokenization, namely unigram or \( n\)-gram(1,1), bigram or \( n\)-gram(2,2), trigram or \( n\)-gram(3,3), \( n\)-gram(1,2), \( n\)-gram(1,3), \( n\)-gram(1,4), \( n\)-gram(2,3), \( n\)-gram(2,4) and \( n\)-gram(3,4). For instance, \( n\)-gram(1.3) is the combination of several types of feature tokenization with a minimum N value of 1 and N maximum of 3. Feature tokenization is then combined with four types of feature weighting, namely TF-IDF, TF-RF, TF-OR, and TF-CHI. The comparison of accuracy results are shown in Figure 3.

![Figure 3](image)

**Figure 3.** Comparison of accuracy in different feature tokenization schemes.

Based on Figure 3, there is a decrease in the accuracy value if the combination of feature tokens does not involve the tokenization of the unigram feature. The highest accuracy is 85%, achieved in the TF-IDF with \( n\)-gram(1,3) and in TF-OR with \( n\)-gram(1,1). The average of accuracy is shown in Figure 4.
As seen in Figure 4, the highest accuracy mean is the TF-IDF (79.8%). The lowest accuracy mean is the TF-CHI (76.7%). In addition to accuracy, the performance is also measured by its precision, shown in Figure 5.

As shown in Figure 5, the highest precision is 92%, achieved in the TF-IDF with the n-gram(2,2). Other than TF-IDF, optimal precision is also seen in n-gram(2,2), which is 90.8% for TF-RF, 89.8% for TF-OR and 89.7% for TF-CHI. The average of precision is depicted in Figure 6.
Figure 6. Precision mean.

The highest precision mean is achieved in the TF-IDF (83.2%), while the lowest precision mean is 80.8%, achieved in the TF-CHI weighting feature. Performance comparison using recall as the metric is shown in Figure 7.

Figure 7. Comparison of recall in different feature tokenization schemes.

The recall is decreased in the TF-IDF with n-gram(2,2) but increased to 98.2% with the n-gram(3,3) and n-gram(3,4). This also occurs in the TF-RF, TF-OR and TF-CHI, with a recall of 98.2% with n-gram(3,3) and n-gram(3,4). As for the average of recall, shown in Figure 8, the highest recall mean is achieved in the TF-IDF, while the lowest is achieved in TF-CHI. The high value of recall (98.2%) does not actually indicate a good classification model as the precision value at 98.2% of recall is low. That way, it is necessary to calculate the harmonic mean of precision and recall represented by the F1-score in Figure 9 and Figure 10.
In Figure 9, the TF-IDF and TF-OR have a maximum F1-score in the n-gram(1,1), while the F1 score of TF-RF is increased with n-gram(1,2). The TF-CHI has an optimal recall value of 80.6% with the n-gram(1,3). The highest F1-score average is achieved on the TF-IDF (82.2%) and the lowest is TF-CHI (78.8%). The F1-score mean is shown in Figure 10.

![Figure 8. Recall mean.](image)

![Figure 9. Comparison of F1 score in different feature tokenization schemes.](image)
5. Conclusion and Future Work
Implementing n-gram feature tokenization and feature weighting can affect the performance of SVM. Based on the average performance of all n-gram combinations, TF-IDF has the highest performance of 79.3% for accuracy, 83.2% for precision, 83.6% for recall and 82.2% for F1-score. The implementation of unigram feature tokenization give an effect to a better performance, measured in accuracy, precision, and F1 score, compared to those with non-unigram feature tokenization.

As for the future work, the preprocessing process may be improved to tackle unsolved word abbreviations in the present study. The future work may also be focused to on-line classification fashion which provide a realtime sentiment analysis and visualized accordingly to portray user satisfaction of Jakarta BRT.

6. References
[1] Badan Pusat Statistik Provinsi DKI Jakarta, “Statistik Daerah Provinsi DKI Jakarta 2015.” Badan Pusat Statistik Provinsi DKI Jakarta, 2015.
[2] “Agustus 2016 Penumpang Transjakarta 11,5 juta.”.
[3] “Global Social Media Statistics Summary 2017.” [Online]. Available: http://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/. [Accessed: 26-Jul-2017].
[4] N. Rahmawati, “Optimalisasi Jumlah Kata Per Tweet Pada Data Latih Dalam Pengembangan Aplikasi Monitoring Kualitas Jaringan Telekomunikasi Seluler Berdasarkan Analisis Sentimen,” Universitas Gadjah Mada, 2017.
[5] A. Tripathy, A. Agrawal, and S. K. Rath, “Classification of sentiment reviews using n-gram machine learning approach,” Expert Syst. Appl., vol. 57, pp. 117–126, 2016.
[6] A. Taukirrahman, “Perbandingan pembobotan fitur pada klasifikasi email spam.”
[7] M. Yusuf and D. D. Santika, “Analisis Sentimen Pada Dokumen Berbahasa Indonesia Dengan Pendekatan Support Vector Machine,” 2011.
[8] P. P. I. Langi, “Pengaruh Stopwords Berdasarkan Chi-Square Pada Sistem Analisis Sentimen Operator Telekomunikasi Seluler Di Indonesia.” Universitas Gadjah Mada, 2016.
[9] B. Pang, L. Lee, H. Rd, and S. Jose, “Thumbs up ? Sentiment Classification using Machine Learning Techniques,” 1988.
[10] M. Lan, C. L. Tan, J. Su, and Y. Lu, “Supervised and traditional term weighting methods for automatic text categorization,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 4, pp. 721–735, 2009.
[11] T. Joachims, C. Nedellec, and C. Rouveirol, “Text categorization with support vector machines: learning with many relevant,” in *10th European Conference on Machine Learning*, 1998.

[12] “Sastrawi 1.0.1: Python Package Index.” [Online]. Available: https://pypi.python.org/pypi/Sastrawi/1.0.1. [Accessed: 17-Jul-2017].

[13] F. Z. Tala, “A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia.”