Sentiment Analysis of Mass Rapid Transit Jakarta using Naïve Bayes Classifier and Rule-based Opinion Target Detection on Twitter

Dhanika Jeihan Aguinta
Faculty of Computer Science
Brawijaya University
Malang, Indonesia
dhanikabin@gmail.com

Putra Pandu Adikara
Faculty of Computer Science
Brawijaya University
Malang, Indonesia
adikara.putra@ub.ac.id

Randy Cahya Wihandika
Faculty of Computer Science
Brawijaya University
Malang, Indonesia
rendicahya@ub.ac.id

ABSTRACT
Jakarta Mass Rapid Transit (MRT) is a transportation that uses electric trains in Jakarta. According to the information on the official website of PT MRT Jakarta, it is known that the MRT Project is a national project that is confirmed by the president of Indonesia in the year 2005 [3]. The presence of Jakarta MRT certainly invites a lot of people’s attention, because MRT is a new facility, especially in the capital city of Indonesia, Jakarta. According to the governor of DKI Jakarta, Anies Baswedan, MRT is not only public transportation but also as culture transformation [6]. Phase 1 MRT lane has been started to operate in 2019 and will be developed further.

1 Introduction
Jakarta, as the center of business, culture, and politics, seeks to increase the quality of comfort for the citizen of the city as well as the comers. The development of Jakarta city can be felt in the modern era like today. One of the aspects that support the development of technology is transportation that is using electrical energy such as electric bicycle and electric car that has started commonly used in this metropolitan city. Not only private vehicles or rented vehicles, but electric buses also started developed to substitute the Trans Jakarta bus as public transportation. The public transportation service in Jakarta is now being developed to be more modern, where one of them is Mass Rapid Transit (MRT).

Jakarta Mass Rapid Transit (MRT) is a transportation that uses electric trains in Jakarta. According to the information on the official website of PT MRT Jakarta, it is known that the MRT Project is a national project that is confirmed by the president of Indonesia in the year 2005 [3]. The presence of Jakarta MRT certainly invites a lot of people’s attention, because MRT is a new facility, especially in the capital city of Indonesia, Jakarta. According to the governor of DKI Jakarta, Anies Baswedan, MRT is not only public transportation but also as culture transformation [6]. Phase 1 MRT lane has been started to operate in 2019 and will be developed further.

As the establishment of phase 1 MRT, indeed there are many people’s opinions regarding that matter. The enthusiasm of the public for the presence of Jakarta MRT can be seen from various social media, which one of them is Twitter. Twitter is social network and also information network that its users can post a
short message as many as 280 characters in one tweet and can be disseminated publicly. Various opinions of the public that is posted on Twitter are not entirely contained positive opinion or negative. The public’s opinion regarding Jakarta MRT can be used as an evaluation for PT MRT Jakarta and the government on the further development of the MRT. Evaluation is intended to consider the policy owner [4].

Things that can be done to know the public’s opinion concerning Jakarta MRT is by having sentiment analysis research. Opinions that have been obtained hereinafter classified into positive or negative using the Naïve Bayes method as text classification. Naive Bayes method is chosen because this method functions well as a text classification method [10]. Besides that, this classification algorithm is modest with the same performance as other algorithms [2].

This research is done to gain things that are significant from the tweet regarding the public’s opinion about Jakarta MRT. The aim of this sentiment analysis is to know and group the public sentiment towards Jakarta MRT. Besides that, it can be used to determine the target object in the opinion regarding the Jakarta MRT, thus from the processed data, the things that need further attention for further development can be identified.

2 Data and Methodology

In this research, the data is gathered from Twitter with small subset of data with total 250 balanced data. After the data is gathered, we annotate manually the tweet into positive and negative tweets. We are using small dataset because this is actually preliminary experiment, since there are already extensive available research using different datasets. The implementation from the algorithm consists of classification calculation on sentiment analysis and rule-based target opinion detection. Several processes are done in the system, namely pre-processing, feature extraction, and finally Naïve Bayes classification, and target detection. The features used in this research are bag of words (BoW), POS tagging using NLTK library, and lexicon-based features. The target detection is employed with chunking that has pre-determined rules. Figure 1 describes the stages of the algorithm implementation.

2.1 Pre-processing

The pre-processing methods used in this research are term normalization by using Prosa.ai API, pre-processing (data cleaning, case folding, tokenization). Term normalization is used to normalize non-standard words (NSW) in Bahasa Indonesia, which commonly occur in social media, into particular standard words. Data cleaning is used to remove unnecessary characters or symbols. Even though characters and symbols such as punctuation or emoticon/emoji can be used as part of textual features, the usage of these features is minimal, thus we decide not to use these features. Case folding is used to change the capitalized letters into lowercase, so that the same word with different capitalization will be considered as one same term. Tokenization is the important step before extracting the features. A sentence is tokenized to produce a list of tokens or terms, which later will be used to extract bag of words features. This list of tokens commonly also known as vocabulary.

2.2 Feature Extraction

Feature extractions consist of several features to help the classification process. The feature that is used to determine the sentiment result is included in the category of lexicon-based features and bag of words. Sentiment selection that is included in negative or positive sentiment is using SentiWords modification [9]. Feature extractions consist of several features to help modelling the classification process. The feature that is used to determine the sentiment result is included in the category of lexicon-based features and bag of words. Sentiment selection that is included in negative or positive sentiment is using modification of SentiWords [9]. The BoW features are extracted from the opinion after going through the pre-processing stage. The lexicon-based features are extracted based on previously proposed lexicon features [8], even though not all are used in this study.

Figure 1: General overview of the system
2.3 Naïve Bayes Classifier

The posterior calculation can be seen in Equation 1 [7]. Naïve Bayes classifier method is one of the linear classification methods with a machine learning approach. The Naïve Bayes algorithm uses training data to gain probability. According to [2], this classification algorithm is modest with the same performance as other algorithms. The posterior calculation is formulated in Equation (1) [7].

\[ P(c|x) = \frac{P(c|x)p(x|c)}{p(x)} \]  

where \( P(c|x) \) is the posterior, the conditional probability of \( x \) in category \( c \), \( P(c) \) is the prior, the document probability with category \( c \), \( P(x|c) \) is the likelihood, the word probability with category \( c \), \( P(x) \) is the evidence, the word probability.

Evidence probability can be eliminated because it has the same probability for all given categories and does not have significant influence on the final result, therefore the calculation is simplified to Equation (2) [7].

\[ P(c|x) = P(c) \times P(x_1|c) \times P(x_2|c) \times P(x_3|c) \times \ldots \times P(x_n|c) \]  

The prior calculation is formulated in Equation 3 [7].

\[ P(c) = \frac{N_c}{N} \]  

where \( N_c \) is the number of documents that belong to category \( c \), and \( N \) is the total number of documents.

The likelihood calculation for Multinomial Naïve-Bayes is formulated in Equation (4) and using add-one/Laplace Smoothing with \( V \) is the number of words in the corpus/training data (vocabulary) to avoid zero probability [7].

\[ P(x|c) = \frac{\text{count}(x,c) + 1}{(\sum_{x \in V} \text{count}(x,c)) + |V|} \]  

where \( \text{count}(x,c) \) is the number of the words in the training document which belong to the category \( c \), \( \sum_{x \in V} \text{count}(x,c) \) is the number of all the word in document training which belongs to category \( c \), and \( V \) is the total number of words in the training document.

2.4 Part-of-speech Tagging

The process of giving a specific tag (word class) in a word is called part-of-speech tagging (POS-tagging). In the process of POS tagging, a sentence is tokenized into words, and each word is determined by each word class using POS tagging methods. The training process and tag determination use the CRFTagger in NLTK POS-Tag library. The POS tagset use in this research is based on previous tagset [5] and the model is trained using Bahasa Indonesia corpus [1].

2.5 Rule-based Opinion Target Detection

To determine the object that becomes the opinion target, we simply apply pre-defined rules over POS Tagging process result. The word structure (syntactic) in a sentence that determines the object in which later becomes the opinion target is identified from the result analysis of POS tagging by using linguistic rules [7]. The necessary rules to detect the object are designed based on word classes that have been previously identified. By using the known word class, chunking step is employed to recognize the entity. Chunking is done by using the NLTK library. Tags that are used to get the opinion target from a tweet is the NP tags that contains entity information. These rules are designed according to the needs of the research object. Table 2 lists the rules to extract the opinion target.

4 Result and Discussion

The testing is done towards the proposed system to evaluate the sentiment analysis and opinion target detection over Jakarta Mass Rapid Transit (MRT) corpus using Naïve Bayes Classifier and rule-based detection. The testing process has two main objectives, to evaluate the effect of lexicon-based features in classification and to evaluate opinion target detection. We use

<table>
<thead>
<tr>
<th>Table 1: Lexicon-based features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature</td>
</tr>
<tr>
<td>F1</td>
</tr>
<tr>
<td>F2</td>
</tr>
<tr>
<td>F3</td>
</tr>
<tr>
<td>F4</td>
</tr>
<tr>
<td>F5</td>
</tr>
<tr>
<td>F6</td>
</tr>
<tr>
<td>F7</td>
</tr>
<tr>
<td>F8</td>
</tr>
<tr>
<td>F9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2: NP tags rules to extract opinion target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
</tbody>
</table>

Example
standard metrics such as precision, recall, and f-measure to evaluate the system.

4.1 The Usage of Features Experiment

The purpose of this experiment is to know which features perform the best. In this experiment, we compare the performance over bag of words (BoW) features, lexicon-based features, and the combination of both features. The result metric using precision, recall, and f-measure of this experiment is shown in Table 3.

Table 3: Performance comparison of feature usage

<table>
<thead>
<tr>
<th>Metric</th>
<th>BoW</th>
<th>Lexicon-based</th>
<th>BoW &amp; Lexicon-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.89</td>
<td>0.88</td>
<td>0.92</td>
</tr>
<tr>
<td>Recall</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>F-Measure</td>
<td>0.94</td>
<td>0.93</td>
<td>0.95</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.89</td>
<td>0.88</td>
<td>0.92</td>
</tr>
</tbody>
</table>

The combination of both features results in the best performance score. According to the Table 3 that shows the effect of the features, it can be seen the combination of BoW and Lexicon-based features achieve the best performance than BoW and Lexicon-based features independently. The usage of BoW features independently lowering the performance because it is also counting the words that have already exist in the training data without knowing the sentiment, while lexicon-based only utilize lexicon or words that have sentiment ignoring the probable words that have sentiment but not identified correctly such as out of scope from the lexicon, incorrect POS Tagging. Hence, the results from the combination of both features could gain the highest performance since it can provide features both semantically and syntactically.

4.2 The Classification and Target Detection Experiment

Classification and target detection testing are done using a confusion matrix to calculate the level of performance from the feature that results in the highest score. The level of performance that is calculated from the confusion matrix includes f-measure, precision, recall, and accuracy. Evaluation score on both classes, which is positive and negative, is calculated by using macro average.

The evaluation score with macro averaging is computed from the evaluation average result of positive class and negative class. The result of macro average from the results can be seen in Table 4. From the obtained result, the ability of the system to classify the negative class is better compared to classifying the positive class. Performance score includes f-measure, precision, recall, and accuracy for negative class get the highest score. Recall score for each positive class and negative get score 1 which means system able to obtain all documents with the correct class. The precision score is 0.79, it happens because it is affected by the number of documents that actually classified as positive class but predicted as negative class. The f-measure score is affected by each precision score and recall. The accuracy with macro average which is on 0.92, with the accuracy in the positive class is lower than the negative class. It happens because it is affected by the incompatibility of several prediction result data with the actual score on positive class, while all of the negative class prediction results are the same with the actual score. All of the performance scores that are obtained in positive class have a lower score than negative class, it happens because it affected by the number of all words in a negative document than there are in all the words in a positive document.

The value of macro average evaluation for opinion target detection can be seen in Table 5. From the precision score, it can be concluded that the system can detect a relevant object correctly by 0.78 from the total of all documents. For recall, the score is 0.85, higher than the precision score. The f-measure score is a score that is balanced weight of precision and recall with having a score of 0.79 and the accuracy is 0.75. Even though the gained result is quite good, but there are still some words
that do not contain information that is still being drawn by the system. Example result of target detection on test data can be seen in Table 6.

The precision score is affected by the obtained of unimportant words. For example, in one of the tweets, there is a word “nama” that do not have important information towards the tweet content. For recall, it is affected by the existence of several words that cannot be tagged correctly, thus those words cannot be drawn by the system in the chunking process. For example, in Table 7 there is the phrase “dukuh atas” that has important information about a place (train station) regarding the object that is pointed in the tweet, but it doesn’t get taken by the system. It happens because it is affected by mistake occurring in the process of POS tagging for the word “dukuh atas” with the NNU tag that causes those two words is not drawn by the system. It is actually can be fixed by revising the rules and need more data for observation.

5 Conclusion

In prototyping sentiment classification and target detection with Naïve Bayes Classifier and Rule-based, the classification process is using Naïve Bayes and combination feature between Bag of Words (BoW) and lexicon-based features. Classification results are classified into negative and positive class. POS Tag chunking rule-based is applied for predicting the opinion target. Evaluation results obtained from Naïve Bayes Classifier methods in classifying the sentiment are precision 0.92, recall 1.0, f-measure 0.95, and accuracy 0.92. The system yields better result when lexicon-based features is combined with Bag of Words (BoW) features. The usage of chunking for target detection resulting in f-measure of 0.79, precision 0.78, recall 0.85, and accuracy 0.75. The POS tagging process for each classification process and target detection has an important influence on the obtained result. The pre-defined rules can also still be further improved to get the correct target.

REFERENCES