Comparison of Naive Bayes and K-nearest neighbours for online transportation using sentiment analysis in social media

A R Atmadja1,*, W Uriawan1, F Pritisen1, D S Maylawati1 and A Arbain2

1 Department of Informatics, UIN Sunan Gunung Djati Bandung, Jalan A H Nasution 105 Bandung, Indonesia
2 Universitas Widya Gama Mahakam Samarinda, Jl. Wahid Hasyim 2 No.28, Kota Samarinda, Indonesia

*aldy@if.uinsgd.ac.id

Abstract. Nowadays, online transportation is one of the transportation that is increasingly preferred by people. It becomes important because people need transportation to be more effective and efficient. However, sentiment analysis is necessary to improve the quality of services on online transportation. Sentiment analysis includes the process of extracting opinions, sentiments, evaluations, and emotions of people about online transportation services on Twitter social media. To get more accuracy in classification, the opinion is taken in large amounts and classify into positive and negative class. There are several steps that use sentiment analysis. Data collection, pre-processing data, POS Tagging, and opinion classification use the Naive Bayes Classifier method, compared to the accuracy of the K-Nearest Neighbours method. The results of the comparison of Naive Bayes Classifier and K-Nearest Neighbours algorithms use 565 data tweets from Twitter, divided 500 trained data, and 65 test data. The result showed that the Naive Bayes Classifier algorithm had achieved the accuracy rate of 66.15%, and K-Nearest Neighbours algorithm produces the accuracy rate of 67.69%. From the results, the K-Nearest Neighbours algorithm perform better accuracy in sentiment classification than Naive Bayes Classifier.

1. Introduction

Transportation is the activity of moving people from one location to another location [1]. According to modern technology, there are several transportation modes such as motorcycle and taxis based on a mobile application. The effect of online transportation mode gives people opinion on social media, such as Twitter, Facebook, Instagram, etc. Online transportation companies have an official Twitter account to provide up-to-date information about services and tweet opinions given by the public to online transportation services, such as application problems, services, and drivers.

Now, there is much research related to text analysis, information extraction, text mining. Sentiment analysis is one of computational research that focused on identify, extract information from text [2]. The text can be taken from social media include opinions, sentiments, and emotion [3]. Sentiment classification can solve with the use of Naïve Bayes Classifier and K-Nearest Neighbours. Naive Bayes classifiers are one of the probabilistic classifiers based on applying Bayes method with strong (naive) independence assumptions between the features [4]. K-Nearest Neighbours (KNN) is a method that uses supervised learning. KNN classification algorithm uses a majority voting of object k and uses the neighbourhood as the predicted score of new query instance [5].
Many research has been done in sentiment analysis. First, the research is contributed in developing sentiment classification in Indonesian Message. The experiment showed that pre-processing noisy text and using dictionaries (positive sentiment word, negative sentiment word, and question word) may increase the accuracy of sentiment classification in social media messages [6]. Second, sentiment analysis can be used for the tourism domain; the result can help managers or service providers understand what the issues are and where improvements are necessary [7]. Another research also has done in making an experiment in the sentiment analysis of a collection of movie reviews that have been automatically translated to Indonesian. The research used three well-known classification techniques: naive bayes, maximum entropy, and support vector machines employing unigram presence and frequency values as the features [8]. Another research are focused on measuring customer satisfaction of online transportation. The research analyses Twitter data by conducting sentiment analysis to tweets containing keywords about GO-JEK and Grab [9].

Determination of sentiments for a text can be done manually, but it has more time to solve the classification problem. So, it is necessary to apply machine learning methods to classify sentiments of tweet opinions [10]. In this study aims to compare two classification method in sentiment analysis using Naïve Bayes and K-Nearest Neighbours algorithm that focused on the data tweet from Twitter for Online Transportation.

2. Methods
In this study, sentiment classification steps consist of five steps: Data collecting, Data labelling, Data Pre-processing, Sentiment Classification, and Measurement. As describe in the part of the introduction, there are two classifications was tested Naïve, using Bayes Classifier and K-Nearest Neighbours. The stages are shown in the below figures.

**Figure 1.** The process of sentiment classification of tweet.

2.1. Data collecting
The analysis system takes opinion data related to online transportation company services from the Twitter website. The process of retrieving data uses technical scrapping over the internet. After scrapping is done, the tweet data is then stored in the database.

2.2. Data labelling
After collecting the data from Twitter and storing data database. The data processed for data labelling. The process of data labelling was completed manually with selecting data with sentiment label. The label for tweet opinion is negative or positive. It means negative when someone voices tweets of inconvenience in the online transportation. While opinion is positive when someone gets good service from online transportation.
2.3. Data pre-processing
The important step in classifying text for sentiment is data pre-processing. In this research, there are four steps done in pre-processing: cleaning, tokenizing, filtering and stemming. The details for data processing steps are described below.

2.3.1. Cleaning. Cleaning stage is the process of cleaning data from words that are not needed. This step is needed to reduce noise, such as removing hashtag, username, and URL. Besides, the cleaning process also involves case-folding. Case folding is use for converting letters into lowercase letters.

2.3.2. Tokenizing. The second steps, the system performed the tokenizing process. Tokenizing is used to separate sentences in a review into single word chunks. For example, the following are the results of tokenizing steps indicated by words separated by "|" delimiter.

| Sentences                                                                 | Tokenizing process                      |
|---------------------------------------------------------------------------|----------------------------------------|
| kesimpulannya tolak kenaikan tarif angkutan                              | kesimpulannya | tolak | kenaikan | tarif | angkutan |
| selamat atas dinaikannya harga tarif subsidi oleh                        | selamat | atas | dinaikannya | harga | tarif | subsidi |
| perusahaan keputusan yang tepat dan rasional                            | oleh | perusahaan | keputusan | yang | tepat | dan | rasional |

2.3.3. Filtering. After tokenizing steps, the filtering process are done by removing non-essential words or words that frequently appear (using a stopword list). For example: “yang”, “ke”, “dari”, “di”, “oleh”, and etc. Next, the word is simplified into original form by removing conjunctions, pronouns and others.

| Sentences                                                                 | Stemming process                       |
|---------------------------------------------------------------------------|----------------------------------------|
| kesimpulannya | tolak | kenaikan | tarif | angkutan                              | kesimpulan | tolak | naik | tarif | angkut |
| selamat | dinaikannya | harga | tarif | subsidi | selamat | naik | harga | tarif | subsidi | perusahaan | putus | tepat | rasional |
| perusahaan | keputusan | tepat | rasional |

2.3.4. Stemming. Stemming is the process of removing affixes, prefixes, endings which aim to change words according to their basic words. The stemming process are shown as follows.

| Sentences                                                                 | Stemming process                       |
|---------------------------------------------------------------------------|----------------------------------------|
| kesimpulannya | tolak | kenaikan | tarif | angkut |
| selamat | dinaikannya | harga | tarif | subsidi |
| perusahaan | keputusan | tepat | rasional |

2.4. Sentiment classification

2.4.1. Naive Bayes classifier. Naive Bayes is a simple probabilistic algorithm [11]. This technique is called 'naive' because it ignores the possibility of dependence or input variation and reduces multivariate problems to a group of uni-variate problems [12]. The data pre-processing stage produces data that is ready for sentiment classification by using the Naive Bayes Classifier method. The classification process requires labels on each data training. Training data are used to determine positive or negative sentiments. The following is an example of the process of classifying or grouping words using the Naive Bayes Classifier method.

| Sentences                                                                 | Words | Words label | Label Opinion |
|---------------------------------------------------------------------------|-------|-------------|---------------|
| kesimpulanya tolak kenaikan tarif angkutan                               | kesimpulan | Positive | Negative |
| tolak                                                                     | Negative |
| naik                                                                       | Negative |
| tarif                                                                      | Positive |
| angkutan                                                                   | Negative |
Table 3. Cont.

| Sentences                                                                 | Words  | Words label | Label Opinion |
|---------------------------------------------------------------------------|--------|-------------|---------------|
| Selamat kenaikan harga tarif subsidi perusahaan putusan yang tepat dan rasional | selamat | Positive    | Positive      |
|                                                                           | naik   | Negative    |               |
|                                                                           | harga  | Positive    |               |
|                                                                           | tarif  | Positive    |               |
|                                                                           | subsidi | Positive  |               |
|                                                                           | perusahaan | Positive |               |
|                                                                           | putus  | Negative    |               |
|                                                                           | tepat  | Positive    |               |
|                                                                           | rasional | Positive |               |

The data from table 1 is used to calculate the Naive Bayes Classifier using the formula below.

\[
K(KP|OP1) = \frac{[K(OP1)K(P)] + K(KP)}{[K(OP1)K(P)] + K(OP1)K(N) + K(KN)}
\]  

(1)

Explanation:
K = Words
KP = Positive Words
KN = Negative Words
OP = Opinion

2.4.2. K-nearest neighbours. In this research, the comparative is done by comparing with the K-Nearest Neighbours (KNN) algorithm. Unlike the Naive Bayes classifier, the KNN algorithm calculates the weight distance which has the highest similarity of each word data to determine the classification [13]. The determination of word weight in this study is as follows.

- Positive words have weights = 6-10
- Negative words have weights = 1-5
- Unclassified/Neutral words have weights = 0

The calculation process of the KNN Algorithm using the Euclidean Distance formula is presented in the following formula.

\[
D(a, b) = \sum_{k=1}^{d} (a_k - b_k)^2
\]

(2)

Explanation:
D: Distance
b: Opinion Weight
a: Word Weight

The following is the data that will be processed for the calculation of KNN.

Table 4. The process of getting opinion weight.

| Sentences                                                                 | Words  | Word Weight | Opinion Weight |
|---------------------------------------------------------------------------|--------|-------------|----------------|
| kesiimpulanya tolak kenaikan tarif angkutan                              | kesimpulan | 0          | 5              |
|                                                                           | tolak   | 5          |                |
|                                                                           | naik    | 3          |                |
|                                                                           | tarif   | 0          |                |
|                                                                           | angkutan | 4          |                |
| Selamat kenaikan harga tarif subsidi perusahaan putusan yang tepat dan rasional | selamat | 10         | 10             |
|                                                                           | naik    | 7          |                |
|                                                                           | harga   | 0          |                |
|                                                                           | tarif   | 0          |                |
|                                                                           | subsidi | 0          |                |
|                                                                           | perusahaan | 0       |                |
|                                                                           | putus   | 3          |                |
|                                                                           | tepat   | 8          |                |
|                                                                           | rasional | 7          |                |
3. Result and discussion

3.1. Naive Bayes classifier result
The experiment is used 65 tweets using the Naïve Bayes Classifier method with the calculation of the accuracy of Positive and Negative sentiments as follows:

\[
\text{Accuracy NBC} = \frac{(\text{Accurate Total Data Amount})}{(\text{Total Data})} \times 100\%
\]
\[
\text{Accuracy of NBC} = \frac{43}{65} \times 100\%
\]
\[
\text{Accuracy of NBC} = 66.15\%
\]

From the above calculations, the system accuracy is obtained for the calculation of Naïve Bayes Classifier, which is 66.15%. In the below picture, the results of application testing that has been implemented Naïve Bayes Classifier.

![Figure 2. Naive Bayes classifier processing result.](image)

3.2. K-Nearest neighbours result
The experiment is used 65 tweets using the K-Nearest Neighbours method with the calculation of the accuracy of Positive and Negative sentiments as follows:

\[
\text{Accuracy KNN} = \frac{(\text{Accurate Total Data Amount})}{(\text{Total Data})} \times 100\%
\]
\[
\text{Accuracy of KNN} = \frac{44}{65} \times 100\%
\]
\[
\text{Accuracy of KNN} = 67.69\%
\]

From the above calculations, the system accuracy is obtained for the calculation of K-Nearest Neighbours, which is 67.69%. In the below picture, the results of application testing that has been implemented K-Nearest Neighbours.

![Figure 3. K-nearest neighbours processing result.](image)
4. Conclusion
Based on the results, it can be concluded that Naïve Bayes Classifier and K-Nearest Neighbours algorithms can be used to classify sentiment from social media for online transportation. Both of two algorithms produce above 60 percent. The experiment carried out using 565 data tweets which are divided into 500 training data and 65 test data, for the Naïve Bayes Classifier algorithm produces an accuracy rate of 66.15% and for the K-Nearest Neighbours algorithm produces an accuracy rate of 67.69%. However, the K-Nearest Neighbours algorithm performs better accuracy in sentiment classification than Naive Bayes Classifier.

This research can still be further developed by scraping in other social media such as Facebook, YouTube, Instagram, news comment, forum, and etc. In addition, the research can combine spam detection systems to improve the quality of tweet data crawling from Twitter.

Acknowledgments
We would like to say thank you for Research and Publication Centre of UIN Sunan Gunung Djati Bandung that gives the full support for this publication research

References
[1] Abbas S 2000 Manajemen Transportasi (Jakarta: Ghalia Indonesia)
[2] Liu B 2012 Sentiment Analysis and Opinion Mining (Morgan & Claypool Publishers)
[3] Buslim N, Busman B, Sinatrya N S and Kania T S 2018 Analisa Sentimen Menggunakan Data Twitter, Flume, Hive Pada Hadoop dan Java Untuk Deteksi Kemacetan di Jakarta JOIN (Jurnal Online Inform. 3 1 1)
[4] Zulfikar W B and Lukman N 2016 Perbandingan Naive Bayes Classifier dengan Nearest Neighbor untuk Identifikasi Penyakit Mata JOIN (Jurnal Online Inform. 1 2 82–86)
[5] Gerhana Y A, Atmadja A R, Zulfikar W B and Ashanti N 2017 The implementation of K-nearest neighbor algorithm in case-based reasoning model for forming automatic answer identity and searching answer similarity of algorithm case 2017 5th International Conference on Cyber and IT Service Management, CITSM 2017
[6] Aqsaath Rasyid N and Purwarianti A 2011 Sentiment Classification for Indonesian Message in Social Media Int. Conf. Electr. Eng. Informatics 17 2–5
[7] Alaei A R, Becken S and Stantic B 2019 Sentiment analysis in tourism: capitalizing on big data Journal of Travel Research 58 2 175-191
[8] Franky and Manurung R 2008 Machine Learning-based Sentiment Analysis of Automatic Indonesian Translations of English Movie Reviews Int. Conf. Adv. Comput. Intell. Its Appl. 1
[9] Anastasia S and Budi I 2017 Twitter sentiment analysis of online transportation service providers Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2016 359–365
[10] Li Y and Fleyeh H 2018 Twitter Sentiment Analysis of New IKEA Stores Using Machine Learning Int. Conf. Comput. Appl. ICCA 2018 4–11
[11] Sinha S, Saxena K and Joshi N 2019 Sentiment Prediction of Facebook Status Updates of Youngsters Data, Engineering and Applications (Springer Singapore) pp 95–105
[12] Yuliani S Y, Sahib S, Abdollah M F, Al-Mhiqani M N and Atmadja A R 2018 Review study of hoax email characteristic Int. J. Eng. Technol. 7 3 2 Special Issue 2 778–782
[13] Kaushik V K, Gupta A K, Kumar A, Prasad A and Lalitha B 2018 Sentiment Analysis of Event Driven Stock Market Price Prediction J. Netw. Commun. Emerg. Technol. 8 4 120–125