Comparing Classification Algorithm With Genetic Algorithm In Public Transport Analysis

Comparing Classification Algorithm With Genetic Algorithm In Public Transport Analysis

Riska Aryanti[1*], Andi Saryoko[2], Agus Junaidi[3], Siti Marlina[4], Wahyudin[5], and Lia Nurmalia[6]

[1]Ilmu Komputer, Universitas Bina Sarana Informatika, Jakarta, Indonesia
[2]Teknik Informatika, Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri, Jakarta, Indonesia
[3,5]Teknologi Informasi, Unversitas Bina Sarana Informatika, Jakarta, Indonesia
[4]Sistem Informasi, Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri, Jakarta, Indonesia
[6]Bahasa Inggris, Unversitas Bina Sarana Informatika, Jakarta, Indonesia

E-mail: riska.rts@bsi.ac.id

Abstract. Congestion major cities in Indonesia caused by the proliferation of the use of private vehicles. Some people express their opinions and its opinion regarding public transport users through social media sites and other websites. This opinion can be used as a sentiment analysis material to find out whether the public transport service is positive or negative. The results of the sentiment analysis can help in the assessment and evaluation of the use of public transportation, it is also expected to improve services and facilities from public transportation so that the public tends to have a positive opinion. Based on the results of the sentiment analysis, it is expected that the community will switch to using public transportation which will certainly reduce congestion. In this study also added preprocessing stages by using the GataFramework framework to complete processes that cannot be done on RapidMiner tools. The method used in this study is sentiment analysis with the method of applying genetic algorithms for feature selection with comparative classification algorithms. Performed by testing the composition of various data. From the results of testing for the case in this study, it was found that the Support Vector Machine classification algorithm based on Genetic Algorithms had a fairly good average accuracy of 76.11% and AUC value of 0.778% with the Fair Classification diagnosis level compared to the three methods such as Naive Bayes, Support Vector Machine and Naive Bayes based on Genetic Algorithms. So that in this study Support Vector Machine classification algorithm based on Genetic Algorithm can be recommended as an algorithm classification good enough to analyze land transportation public sentiment. Based on the analysis it is expected that the public sentiment will switch to using public transport which would reduce congestion.

1. Introduction

These facts make it imperative to improvise public transportation. The previous research has been done to introduce new technological solutions in the scope of services in mobility to minimize problems arising from big cities such as traffic congestion, greenhouse gas emissions, fuel consumption, and others. Many works propose joint service models to make transportation services more efficient both for business models and a reduction in travel costs for passengers [1]. According to G. Arnould in [2] most shared transportation services in the literature framework

is Carpooling and Taxi, this service is intended to share one vehicle for several passengers in one trip, and it may use or it may not have local searches and similar destinations.

Many countries have adopted a range of actions and policies to develop public transport and have even introduced priority strategies to encourage the development of public transportation. However, many opportunities for development, according to Murray in [3] public transportation also faces many challenges. There is the unavailability of proper and integrated public transportation all around and economically affordable. So that people no longer use private vehicles as daily transportation, which is no less important is the government has not made efforts to increase public awareness of the importance of an orderly and obedient culture of traffic signs and rules. The solution to reducing congestion is to increase the use of public land transportation in the city, which in fact is still not much in demand by the public.

According to A. M. Kaplan in [4] Lately, social media has become an extraordinary trend. The role of social media is very influential for the development of the current global situation. According to Forrester Research, 75% of internet surfers used social media in the second quarter of 2008 by joining social networking sites, reading blogs, or simply giving reviews for online shopping sites, a significant increase of 56% in 2007, this growth is not limited to youth groups because those who are now in the Generation X group (now in the age range of 35-44 years) also participated in it, either limited to just joining, just listening, or critics in it. he language of online reviews is dif-ferent from formal language, which means it not only includes domain-specific words but also a lot of Internet catchwords, making it difficult to create a compre-hensive and accurate domain dictionary [5].

Some people express their opinions about the use of public transportation in the city, the people's opinions are expressed on social media, one of which is Twitter. Twitter is one of the social media that is used to express a review of various issues or topics that are trending through the tweet column. But reading the review as a whole can be time-consuming. Meanwhile, if only a few reviews are read, then the evaluation will be biased. Sentiment analysis aims to overcome this problem by automatically grouping user reviews into positive or negative opinions.

The analysis of sentiment using Indonesian review still has difficulties in the process of preprocessing, which is the process by which data retrieved directly from Twitter should still be Text uniformity for the next process. RapidMiner tools are not yet available in the English dictionary for the preprocessing of text with Indonesian language. Gataframework is A text mining framework in Bahasa Indonesia, Gataframework has a feature for process the text so that a sentence can be preprocessing well, so that with using Gataframework can be a solution to the stage of preprocessing text mining Indonesian language. Analyzing the text computational linguistics are used to deduce and analyze mental knowledge of Web, social media and related references [6]. Sentiment analysis is to capture opinions and emotions in user reviews that can be used to predict customer demands and to provide support in company decision making [7]. Sentiment analysis is an activity carried out to see the level of public sentiment or public opinion relating to goods or services and even a figure, both political and celebrity figures [8].

Sentiment analysis is a kind of text classification that classifies text based on the orientation of the sentiment of opinion it contains. This is also known as opinion mining, opinion extraction, and influencing analysis in the literature [9]. There are several classification algorithms that can be used for text classification sentiment analysis including Naive Bayes (NB), Support Vector Machine (SVM), and K-Nearest neighbour (KNN). SVM can solve the classification problem, but SVM has a weakness in the difficulty of selecting appropriate and optimal features in the weight of the attributes used, causing the classification accuracy to be low. This research starts from a problem in the text classification on Twitter which consists of approximately 140 characters using SVM, where the classification has a lack of problems with the selection of the appropriate parameters, because with a mismatch a parameter setting can cause low classification results [10].

Previous research has been carried out by applying Genetic Algorithm (GA) for feature selection in analyzing sentiments using classification algorithms such as NB, and SVM which are most commonly used. Research conducted by [9] the study Classifier based on Genetic Algorithm in the Text Classification of Railway Fault Hazards use TF-IDF method to extract text features and convert them into vectors Then, decision tree classifier is used to classify the data. In order to improve classification accuracy, the Bagging Ensemble Classifier conduct a random sample training to text vector converted by TF-IDF which decision as the base classifier, produce Bagging classification results, considering the Bagging Algorithm is the number of base classifiers results voting combination classification model which has a better classification performance, we use Genetic Algorithm to calculate Bagging [11].

There are several feature selection techniques that can be used to solve optimization problems, including GA. GA has the potential to produce better features and become optimal parameters at the same time [10]. Thus, this study chose the feature selection technique using GA which will compare one by one to the 2 (two) classification algorithms namely NB, and SVM, both of these algorithms to be applied in classifying text in a review of public land transportation in cities, which the result can be determined from the application of GA for feature selection by comparing which classification algorithm is best to be applied in order to improve the accuracy of sentiment analysis.

2. Methods

The method of research is the experimental research, it has some stages:

2.1. Data Collection Method

The first step in this research is data collection on public transportation user reviews from social media such as Twitter by selecting 4 sample vehicles like as busway, commuter line, motorcycle taxi and taxi. The dataset focuses on Indonesian-language opinions that discuss the use of public land transportation specifically in cities. The next stage is the process of labeling data by giving status tweets based on positive sentiments and negative sentiments. In the labeling process, this is done manually by giving information values for each tweet. The following are examples of data that have been given positive and negative status.

At the initial stage, raw data is collected using the Twitter search operator available on the RapidMiner tool with the help of using the Twitter Access API as a connector to the Twitter API about public transportation users based on the keywords “Busway”, “Commuterline”, “Ojek”, and “Taxi” with data is taken from July 15 to July 29, data collected in the process of crawling tweet data is stored in the excel file format. After the Twitter data is finished to be collected, the next step is to label the tweets according to the predetermined class. Dataset has been collected from Twitter is data that does not have a label (unsupervised data), in order to be processed using supervised learning techniques, then Twitter data that has been collected previously needs to be labelled. In this study, the labelling process is done manually by giving positive and negative tweet status.

2.2. Initial Data Processing

Initial data processing is the preprocessing stage carried out by the process of tokenization, transform cases, stopword removal, normalize, stemming and generate N-grams. At this stage, It use additional Gataframework because the RapidMiner tools still have weaknesses in the Indonesian language text, the writer uses Gataframework http://www.gataframework.com/textmining/ for initial data processing such as: stopword removal, normalize Indonesian slank, and stemming

2.3. Classification

The next stage is the classification process by comparing classification models such as support vector machines, and Naive Bayes was chosen to determine the suitability of the data through method is the best of some of the text classification methods used by some previous researchers. Support Vector Machines (SVM) is a guided learning method that analyze data and recognizes patterns, used for classification and regression analysis. This method was developed by Boser, Guyon, Vapnik, and was first presented in 1992 at the Annual Workshop on Computational Learning Theory. Patterns that are joined in a class - 1 are symbolized in red (box), while patterns in class +1, are symbolized in yellow (circle). The classification problem can be translated as an attempt to find a line (hyperplane) that separates the two groups [12].

Feature selection is an important function in classification and prediction technique, especially in medical data mining. It is embedded with the task of selecting a subset of relevant features that can be used in constructing a model [13]. Feature selection is imperative in machine learning and data mining when we have high-dimensional datasets with redundant, nosy and irrelevant features. The area of feature selection deals reducing the dimensionality of data and selecting only the most relevant features to increase the classification performance and reduce the computational cost [14].

Naive Bayesian is a classification method based on probability, there is assuming that each X variable is independent. In other words, Naïve Bayesian assumes that the existence of an attribute has nothing to do with being in another attribute. If it is known that X is sample data with unknown class (label), H is a hypothesis that X is data with class (label) C, P (H) is the probability of hypothesis H, P is the chance of observed sample data, then P (X — H) is the probability of sample X data, if it is assumed that the hypothesis H is valid (valid) [12].

2.4. Weighting and Selection of Selection Features

The next step is to add a weighting model because the classification has shortcomings to the problem of selecting the appropriate parameters, because it is not matching a parameter setting can cause the classification results to be low. Method of weighting the features to be used are Term Frequency Invers Document Frequency (TF-IDF) and feature selection using Genetic Algorithm (GA). The classification used is first using Naive Bayes and then the second classification using Support Vector Machine with 10 Fold Cross Validation test.

Genetic Algorithm (GA) is one of the first population-based stochastic algorithm proposed in the history. Similar to other EAs, the main operators of GA are selection, crossover, and mutation. This chapter briefly presents this algorithm and applies it to several case studies to observe its performance[15]. Genetic algorithms have the disadvantage that the selection of the wrong parameters can reduce the accuracy produced.

2.5. Validation

The algorithm accuracy results will be illustrated in the Confusion Matrix and ROC curves. RapidMiner is used as a tool in measuring the accuracy of experimental data conducted in research. Validation is the process of evaluating the accuracy of a model. In evaluating classification models based on the calculation of data objects testing which are predicted to be true and incorrect. These calculations will be tabulated into a table called a Confusion Matrix[16].

3. Results And Discussion

The implementation of the research method in data analysis started from data collecting, initial data processing, classification, weighting and selection of selection features, and validation. The results of a comparison of testing experiments using the Naive Bayes model, Support Vector

Machine, Naive Bayes based on Genetic Algorithms and Support Vector Machines based on Genetic Algorithms can be seen in table 1 and table 2.

Table 1. Accuracy Results Comparison of NB, SVM, NB-GA and SVM-GA Algorithms.

Dataset	NB	SVM	NB-GA	SVM-GA
Ojek	77,94	79,97	71,59	81,06
Busway	85,27	85,51	88,55	85,39
Commuter line	71,02	72,98	67,17	74,75
Taxi	60,1	61,25	63,06	63,22
Average	73,58	74,93	72,59	76,11

Table 2. AUC Value Comparison of NB, SVM, NB-GA and SVM-GA Algorithms.

Dataset	NB	SVM	NB-GA	SVM-GA
Ojek	0,526	0,817	0,500	0,783
Busway	0,819	0,953	0,813	0,952
Commuter line	0,479	0,739	0,500	0,697
Taxi	0,458	0,666	0,551	0,680
Average	0,571	0,794	0,591	0,778

Table 1 and Table 2 shows that the Support Vector Machine + Genetic Algorithm has the highest average and the ratio of the AUC with Support Vector Machine algorithms can be shown in Figure 1

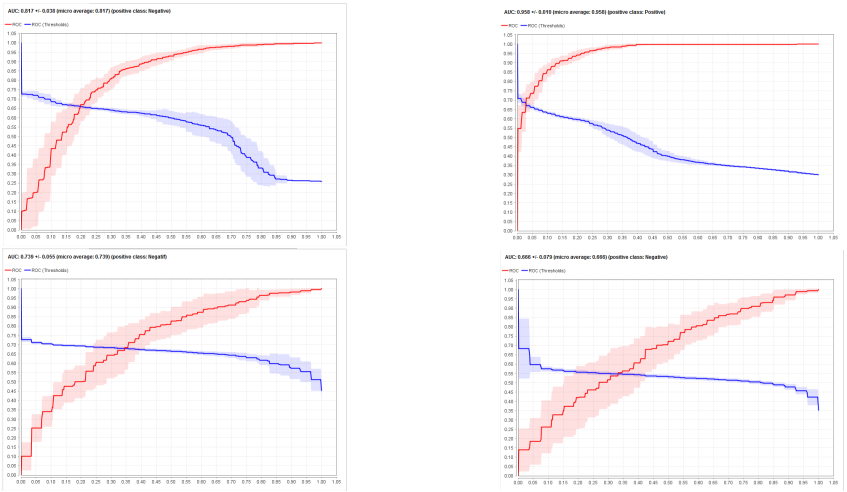

Figure 1. the highest average and the ratio of the Accuracy with Support Vector Machine using Genetic Algorithms

4. Conclusion

The Support Vector Machine algorithm is superior to the Naive Bayes algorithm in classifying public opinion with Indonesian text on Twitter for the use of public land transportation in cities. The average accuracy for the Naive Bayes algorithm reached 73.58% and the AUC value reached 0.571%, while the Support Vector Machine algorithm has the average accuracy reached 74.93% and the AUC value reached 0.794%. After adding the Genetic Algorithm selection feature, the average accuracy for Naive Bayes using Genetic Algorithm is 72.59% and the AUC value is 0.591%, while for the Support Vector Machine algorithm using Genetic Algorithm has the average accuracy results increase by 0.778%, so the average results of Support Vector Machine algorithm using Genetic Algorithms can be recommended as a fairly good classification in sentiment analysis of public transportation. The study also found that using the Gataframework framework could help with the text mining preprocessing stage for Bahasa Indonesia. The Feature in Gataframework makes text drawn from Twitter social media into data that can be processed in RapidMiner tools. This research can also be developed by creating a sentiment analysis system using the SVM-GA classification method to be recognized in real-time.

Acknowledgement

The author would like to thank all those who participated in this research, especially the respondents who filled out the questionnaire in this research.

References

[1] Hosni H Naoum-Sawaya J and Artail H, 2014 The shared-taxi problem: Formulation and solution methods Transp. Res. Part B 70 p. 303–318.
[2] Ulloa D Saleiro P Rossetti R J F and Silva E R, 2016 Mining Social Media for Open Innovation in Transportation Systems in International Conference on Intelligent Transportation Systems (ITSC) p. 169–174.
[3] Li L Bai Y Song Z Chen A and Wu B, 2018 Public transportation competitiveness analysis based on current passenger loyalty Transp. Res. Part A Policy Pract. 113, April p. 213–226.
[4] Kristiayanti D A Umam A H Wahyudi M Amin R and Marlinda L, 2018 Comparison of SVM & Naïve Bayes Algorithm for Sentiment Analysis Toward West Java Governor Candidate Period 2018-2023 Based on Public Opinion on Twitter in International Conference on Cyber and IT Service Management (CITSM 2018) Citsm p. 3–8.
[5] Li R Lin Z Lin H Wang W and Meng D, 2018 Summary of text emotional analysis J. Comput. Res. Dev. p. 55, 30–52.
[6] Kacprzyk J, 2020 Recent Advances in NLP: The Case of Arabic Language Stud. Comput. Intell. 534 p. 1–292.
[7] Meng W Wei Y Liu P Zhu Z and Yin H, 2019 Aspect Based Sentiment Analysis with Feature Enhanced Attention CNN-BiLSTM IEEE Access 7 p. 167240–167249.
[8] Wongkar M and Angdresey A, 2019 Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter Proc. 2019 4th Int. Conf. Informatics Comput. ICIC 2019 p. 1–5.
[9] Govindarajan M, 2013 Sentiment Analysis of Movie Reviews using Hybrid Method of Naive Bayes and Genetic Algorithm Int. J. Adv. Comput. Res. 3, 13 p. 139–145.
[10] Wahyudi M and PUTRI D A, 2016 Algorithm Application Support Vector Machine With Genetic Algorithm Optimization Technique for Selection Features for the Analysis of Sentiment on Twitter J. Theor. Appl. Inf. Technol. 84, 3 p. 321–331.
[11] Li X Shi T Li P and Zhou W, 2019 Application of Bagging Ensemble Classifier based on Genetic Algorithm in the Text Classification of Railway Fault Hazards 2019 2nd Int. Conf. Artif. Intell. Big Data, ICAIBD 2019 p. 286–290.
[12] Wahyuni E S, 2016 Penerapan metode seleksi fitur untuk meningkatkan hasil diagnosis kanker payudara J. SIMETRIS 7, 1 p. 283–294.
[13] Benghozi P J Krob D and Rowe F, 2019 Advances in Intelligent Systems and Computing: Preface Adv. Intell. Syst. Comput. 205 AISC.
[14] Bansal J C Delhi N Deep K and Nagar A K, 2020 Evolutionary Machine Learning Techniques.
[15] Kacprzyk J, 2019 Evolutionary Algorithms and Neural Networks Stud. Comput. Intell. 534 p. 1–292.
[16] Gorunescu F, 2011 Data Mining Concepts, Models and Techniques .