Gojek and Grab User Sentiment Analysis on Google Play Using Naive Bayes Algorithm And Support Vector Machine Based Smote Technique

Gojek and Grab User Sentiment Analysis on Google Play Using Naive Bayes Algorithm And Support Vector Machine Based Smote Technique

Hermanto[1], Antonius Yadi Kuntoro[2], Taufik Asra[3], Eri Bayu Pratama[4], Lasman Effendi[5], and Ridatu Ocanitra[6]

[156]Teknologi Komputer, Universitas Bina Sarana Informatika
[2]Sistem Informasi, STMIK Nusa mandiri
[3]Rekayasa Perangkat Lunak, Universitas Bina Sarana Informatika
[4]Sistem Informasi, Universitas Bina Sarana Informatika

E-mail: hermanto.hmt@bsi.ac.id

Abstract. As the online Ojek services, people often talk about them by giving their opinions and opinions through various media, one of which is Google Play opinion given by the public to the services of online Ojek also diverse. Users provide review reviews or comments about the application, of course users will choose an app that has a good review. But monitoring the reviews of the general public is not easy, because the amount is very much to be processed so that researchers want to know the extent of the user review analysis of Gojek and Grab applications based on the review of user comments using the classification technique is using the NB algorithm and SVM based technique Smote. The results of the test with the highest accuracy result 81.09% and AUC value = 0.922 is the application Gojek while for application test results grab accuracy value of 73.20% and AUC value = 0.848. To that end, the implementation of the Support Vector Machine based Smote technique in this study has higher accuracy so that it can be used to provide solution to the sentiment analysis problems in the review user comments online Ojek application

1. Introduction

Transportation is one of the supporting needs of people who use as a means to move from one place to another. Along with the development of technology today contribute to the performance of systems in all aspects included in the transportation aspect. Nowadays, Ojek online is becoming the latest public transportation trend among the community because of the ease in using this transportation service through the application device without having to conventionally come to conventional place to use this transportation service. As the online Ojek services, people often talk about them by giving their opinions and opinions through various media, one of which is Google Play opinion given by the public to the services of online Ojek also diverse. , books, apps, games, or media players. Google Play can be accessed through the Web, Android apps (Play Store), and Google TV. In Google Play, in addition to the online products store also includes a valuation feature for customers to give reviews of the advantages and disadvantages in the use of online Ojek applications, user reviews can be various forms, ranging from subtle and rough sentences also exist depending on the assessment of each user. Google Play users, in

addition to providing reviews, can also provide ratings in the form of a star score (between 1 – 5) depending on the rating given by the user and the last user can also give a "like" score to someone who can indeed represent someone's interest or dislike of the online Ojek application. For this research, the main focus is on the Gojek and Grab app user reviews on Google Play website. Gojek or GrabTaxi user reviews can be influenced by some of the things that are not yet a concern from either Gojek or Grab. This may be due to several factors that have to be fixed and not yet known by Gojek or Grab. Opinions given by the public about the services provided by the online Ojek services are varied such as providing opinions on the satisfaction or decrease of public satisfaction using this service, so with the number of opinions given, make the community become selective in selecting the online Ojek service provider with the condition, the company side of the service provider of online Ojek can also know the public satisfaction of the services provided so as to improve the continued increase the quality of services that will be provided to the customer. From the set of backgrounds outlined above, researchers will perform Gojek and Grab user Review analysis based on users ' comments online Ojek using Naive Bayes and SVM based Smote methods. From the process of testing data on the RapidMiner began with the formation of models with the data in the first part of data sharing training and data testing. After testing the data generated Accuracy by looking at the confusion matrix and performance measured using accuracy and AUC and displayed in the form of ROC curves, certainly can be known application of online Ojek which has the highest level of accuracy according to the results of the research using the algorithm model Naive Bayes and SVM based technique Smote.

The sentiment classification against the reviews available online include, Comparison of Naïve Bayes Algorithm, C 4.5 and Random Forest for Service Classification Ojek Online [7].

Facebook Social Media Comments Data sentiment analysis with Knearest Neighbor (case study on freight forwarding service account J&T Indonesia) [1].

The study proposed comparisons between the four most popular classification algorithms namely Naive Bayes, SVM, Decission Tree and Random Forest algorithms in Amazon Unlocked Mobile review datasets. The results showed that SVM was the best algorithms by achieving the highest score in all metrics both accuracy, precision, recall and F1 score [12].

In 2019 this study conducted a grouping of opinions from college alumni using SVM and NBC algorithms in which the study was divided into 3 main components i.e. input components, opinion grouping systems, as well as components of output against positive opinions, neutral opinions and negative opinions. The results of the study's highest accuracy value on the NBC algorithm reached 94.45% while the highest level of accuracy in the SVM algorithm reached 75.76% [13].

2. Method

Sentiment analysis is a process for determining the sentiment or opinions of someone embodied in text and can be categorized as a positor negative sentiment [4]. Sentiment analysis refers to a broad field of natural language processing, linguistic computing and text mining aimed at analyzing the opinions, sentiments, evaluations, attitudes, judgments and emotions of one's whether the speaker or the author is concerned with a particular topic, product, service, organization, individual, or activity [9]. Sentiment analysis is also a computational research of opinions, sentiment and textual emotions expressed [8].

2.1. Text Mining

The important purpose of text mining is to get high-quality information from text. This is usually done through the discovery of patterns and trends in ways such as statistical pattern learning, topic modeling, and statistical language modeling. Text mining usually requires structuring input text (for example parsing, along with the addition of some inherited linguistic features and patterns taken from a structured database). This is followed by lowering patterns

in structured data, and evaluating and interpreting outputs. In text mining is usually able to analyze semi-structured text data referring to a combination of relevance, novelty, and interest [5]. Text mining is a part of data mining where the process is mainly carried out by extracting knowledge and information from patterns contained in text form using certain analysis tools [6].

2.2. Support Vector Machine (SVM)

SVM is a learning machine method that works on the principle of Structural Risk Minimization (SRM) with the aim of finding the best hyperplane that separates two classes in the input space. In short, an SVM looks for the best hyperplane that functions as a separator of two classes of data. In this new dimension, it seeks a hyperplane separating linear optics (i.e., "decision limit " Separates tuples from one class of another). SVM attempted to find Hyperplan using vector support ("essential " training tuples) and margins (determined by vector support) [6].

2.3. Naïve Bayes

The Naive Bayes algorithm is a statistical classifications that can be used to predict the probability of a class membership. According to Wu and Kumar that Naive Bayes was a popular classification method and entered in the top ten algorithms in data mining. Naive Bayes uses a mathematical branch known as probability theory to find the greatest opportunity of classification, by looking at the frequency of each classification on the training data [7].

2.4. Teknik SMOTE

The SMOTE (synthetic Minority Over-sampling Technique) oversampling method used to deal with class imbalance problems. This technique synthesizes new samples from minority classes to balance the dataset by creating new instances of minority classes by forming convex combinations from neighboring instances. Then, the required oversampling level is chosen randomly. Effectively draws a line between the minority points in the feature space and the sample along this line. Examples of this new minority are added to training data, and classifiers are trained with additional data. The SMOTE algorithm is generally more accurate than the usual oversampling approach [3].

In this study, Gojek and Grab user reviews took data from Google Play, where user reviews were taken a total of 2160 data, for Gojek user reviews amounting to 1160 and Grab user reviews amounting to 1000. To test the dataset with naïve Bayes and smote-based SVM result of testing both algorithms we will compare the accuracy value and the AUC to be implemented into an information system to classify the user reviews Gojek and grab against the comments whether positive or negative.

3. Result and Discussion

At the stage of bussines understanding, conducted with an understanding of the research object. The Data that will be used in this research is a review of the users of Gojek app and grab on Google Play that comments, to determine the status of comments on users of Gojek or grab that you have on Google. Motivation at this stage data review user comments Gojek or grab is presented in the form of text on Google Play that will be grouped by the content of the discussion of each review category comments. At this stage, there is also an understanding to find the best method of algorithm in order to petrify during the process of data processing to be done by comparing the results of Naive Bayes and Support Vector Machine (SVM) algorithms on Gojek and Grab's user review comments.

3.1. Data understanding

This is the process of understanding the data that will be used as material to be researched to be done to the following stage, namely processing. Gojek and grab user review data fetches

data from Google Play, where a review of user comments taken a total of 2,160 data, for reviews of Gojek user comments amounting to 1160 and reviews of grab user comments amounting to 1000. From the user's review, it is grouped by the comment categories given by Gojek or grab users

3.2. **Data Preparation**

The next step is to prepare the data that has been obtained so that it can be processed when doing modeling. The preprocessing stage includes the activities of Tokenization, case folding, Filtering, Stopwords removal, and Stemming to be ready to be managed to the next stage. The Preprocessing process in this study uses two preprocessing applications, first using the Gata Framework accessed via the link http://gataframework.com/textmining which can be used free of charge is also easy to use because there is no need to create an account to use the service and continue the RapidMiner preprocess. From the results of pre-processing using the Gata Framework, the data set will be pre-processed again by using the RapidMiner tool to clean the data for better results.

3.3. **Remove Duplicates**

Remove Duplicates is the next step of data preparation used in RapidMiner software. Remove duplicates are used to remove the same or duplicate text. This is done so that the data is not fulfilled by the same text so that it slows the process of running software to analyze the model.

3.4. **Nominal to Text**

Nominal to text is an operator in the RapidMiner that serves to change all the numbers in the text into a text. So that the existing number will be considered type of text data is not numeric or nominal.

3.5. **Transform Case**

The Operator used at this stage is to change the existing capital letters of the Texa to be converted to all lowercase letters. This is done so that the type of process into the classification model is the uniformity of the letter and no error occurs in the tokenize process.

3.6. **Filter Token (By Length)**

Filter Token is a process that exists in the data preparation to remove any number of words (after tokenize process) with a specific character length. In this study the minimum length of character used is 4 characters and a maximum length of 25 characters. This means that a word whose length is less than 4 characters and more than 25 characters will be eliminated.

3.7. **Filter Stopword (Dictionary)**

Next is the use of Stopword Removal (by Directory) operator that serves to eliminate words that are not linked to the text content. In the previous stage using service text mining Gata framework has been done but there are some words that cannot be eliminated by the previous service because it has not been entered as a word to be deleted. Then with the operator Stopword Removal (by Directory) Researchers can register the word that should be removed from the text.

3.8. **Modeling Stages**

It is the stage of selecting mining techniques by determining the algorithm to be used. The results of the model testing carried out were to classify comments from Gojek user comments and retrieve whether the comments were Positive or Negative using the NB and SVM algorithm to find the highest accuracy value.

3.9. **Testing Stages**

The arrangement and use of the operators and parameters in the Rapid Miner frameworks have an effect on the accuracy and model formed, more clearly testing of the three algorithms is as follows:

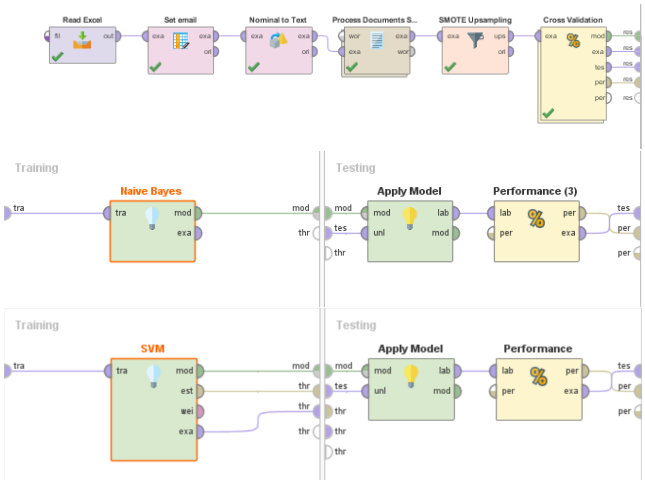

Figure 1. Testing Stages

Figure 2 above is the NB algorithm testing model and SVM with Smote using rapidminer, starting from entering data then managing the set of roles that will determine the label there and the nominal text and then process the document. In this test, the data used is clean data that has been through preprocessing. Data is retrieved using the Read Excel operator which is stored in Excel (.xlsx) format. Then the algorithm design model is processed into the Naive Bayes cross-validation operator and SVM in which there is an algorithm calculation then the model is applied after it is entered into the performance appraisal then the accuracy and AUC results appear.

3.10. **Accuracy and AUC value**

Based on the results of the experiments conducted using the user review data Gojek app and grab on Google Play that gives comments, where in this experiment using the NB algorithm and SVM by using 2160 comment data that has been in the filter then generated Accuracy and AUC values as follows:

Table 1. Accuracy and AUC

	Accuracy	AUC
`Gojek Naive Bayes`	74,41%	0.720
`Gojek SVM + Smote`	**81,09%**	0.922
`Grab Naive Bayes + Smote`	64,93%	0.587
`Grab SVM + Smote`	**73,20%**	0.848

From the results of the performance comparison of the two algorithms above, the test results of support vector Machine better the value of accuracy than the naive Bayes algorithm. Thus, the application of technical-based Support Vector Machine in this study has higher accuracy so it can be used to provide solution to the sentiment analysis problem in the review user comments online Ojek application.

4. Conclusion

In this study, after preprocessing and testing the model using data mining methods, namely Naive Bayes and support vector machines based on smote techniques. It can be seen that the accuracy value to determine that the review analysis of online motorcycle taxi user reviews namely Gojek and Grab, can be proven by the accuracy value and the AUC value of each name for the Gojek application using SVM algorithm based on the smote technique accuracy value = 81.09% and the AUC value = 0.922, while for the Grab application using the SVM algorithm based on the smote technique the accuracy value = 73.20% and the AUC value = 0.848. In the research, it can be seen that the level of accuracy obtained by the Gojek and Grab application uses the Naive Bayes algorithm and Support Vector Machine based on the smote technique, the superior algorithm that has the highest accuracy value is the support vector machine algorithm compared to the Naive Bayes algorithm. In research [7] using SVM with other data yielded an accuracy of 69.18%. For this reason, the application of Support Vector Machine based on the smote technique in this research has higher accuracy so that it can be used to provide solutions to sentiment analysis problems in the review of online Ojek application user comments.

References

[1] Hermanto H Mustopa A and Kuntoro A Y, 2020 Algoritma Klasifikasi Naive Bayes Dan Support Vector Machine Dalam Layanan Komplain Mahasiswa JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer) 5, 2 p. 211–220.

[2] Salam A Zeniarja J and Khasanah R S U, 2018 Analisis Sentimen Data Komentar Sosial Media Facebook Dengan K-Nearest Neighbor (Studi Kasus Pada Akun Jasa Ekspedisi Barang J&T Ekpress Indonesia) Pros. SINTAK p. 480–486

[3] Guia M Silva R R and Bernardino J, 2019 Comparison of Naive Bayes, support vector machine, decision trees and random forest on sentiment analysis IC3K 2019 - Proc. 11th Int. Jt. Conf. Knowl. Discov. Knowl. Eng. Knowl. Manag. 1, January p. 525–531.

[4] Dharmendra K Saputra K O and Pramaita I N, 2019 Analisa Sentiment Untuk Opini Alumni Perguruan Tinggi Maj. Ilm. Teknol. Elektro 18, 2.

[5] Basari A S H Hussin B Ananta I G P and Zeniarja J, 2013 Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization Procedia Eng. 53 p. 453–462.

[6] Ipmawati J Kusrini and Taufiq Luthfi E, 2017 Komparasi Teknik Klasifikasi Teks Mining Pada Analisis Sentimen Indones. J. Netw. Secur. 6, 1 p. 28–36.

[7] Hermanto H Kuryanti S J and Khasanah S N, 2019 Comparison of Naïve Bayes Algorithm , C4 . 5 and Random Forest for Service Classification Ojek Online J. Publ. Informatics Eng. Res. 3, 2.

[8] Hadna M S Santosa P I and Winarno W W, 2016 Studi Literatur Tentang Perbandingan Metode Untuk Proses Analisis Sentimen Di Twitter Semin. Nas. Teknol. Inf. dan Komun. 2016, Sentika p. 57–64.

[9] Kamber M Han J and Pei J, 2012 Data Mining: Concepts and Techniques Waltham, MA: Morgan Kaufmann.

[10] Arwan et al., Synthetic Minority Over-sampling Technique (SMOTE) Algorithm For Handling Imbalanced Data – MTI. [Online]. Available: https://mti.binus.ac.id/2018/06/08/synthetic-minority-over-sampling-technique-smote-algorithm-for-handling-imbalanced-data/. [Accessed: 03-Aug-2020].