Twitter Sentiment Analysis on Online Transportation in Indonesia Using Ensemble Stacking

Yahya Setiawan*, Jondri, Widi Astuti

School of Computing, Informatics Study Program, Telkom University, Bandung, Indonesia
Email: 1yahyasetiawan@student.telkomuniversity.ac.id, 2 jondri@telkomuniversity.ac.id, 3 widiwdu@telkomuniversity.ac.id
Correspondence Author Email: jondri@telkomuniversity.ac.id

Abstract—Online transportation is a transportation innovation that has emerged along with the development of online-based applications that provide many features and conveniences. In its development, many users wrote their responses to the application on social media such as twitter. Many opinions and responses are directly conveyed by users of online transportation modes to their official accounts. The responses given by these users are very large and can be used as sentiment analysis on online transportation. However, the analysis process cannot be done manually. Therefore, we need a system that can help analyze user responses on Twitter automatically. In this study, a sentiment analysis system was built for online transportation in Indonesia using the ensemble stacking algorithm, which will simplify and increase the accuracy of the sentiment analysis. Ensemble stacking is a solution for advanced machine learning methods that can improve the performance of the base classifier. The system built on ensemble stacking uses three base classifiers, namely SVM kernel RBF, SVM linear kernel, and logistic regression. The best accuracy result on the gojeck dataset is 88%, and the best F1 score is 87%. Ensemble Stacking which is applied to the research that the author conducted on online transportation sentiment analysis on twitter, obtained better accuracy than the base classifier used.

Kata Kunci: Ensemble Stacking; Machine Learning; Online Transportation; Sentiment Analysis; Twitter

1. INTRODUCTION

Online transportation is a mode of transportation that is quite popular today with various conveniences and features provided by online transportation that attracts many users. With this popularity, there will be many opinions or opinions on services, features, and promos from online transportation provided by online transportation users [1]. While many online transportation services in Indonesia are pretty famous, two online transportation services have quite a lot of users in Indonesia, namely gojek and grab [2]. With the increasing popularity of online transportation, many online transportation companies create social media accounts to notify the latest promotions or innovations of their products or services [3]. One of the popular social media used is Twitter. Twitter is a text-based social media platform where users can write their opinions or opinions on a particular matter or issue [4]. Where social media Twitter is strong and cheap enough to get information on the sentiments of online transportation users from user tweets [5]. Twitter is the most popular social media in the world. According to available data, there are 500 million active users of social media and more than 500 million tweets every day [6]. Many people write their opinions, thoughts, feelings, and protests on Twitter, so Twitter can be used as a data source that can be analyzed [7]. This data source can be used to analyze a person's sentiment on a topic, one of which is customer satisfaction with online transportation [6]. In addition to the quality of information on Twitter, Twitter also provides several types of Application Programming interfaces (API) that third parties can use. The third is to download data from the Twitter platform itself [8].

With the development of social media Twitter, many people write their opinions, inputs, and criticisms. From the increasing number of Twitter users who continue to write their opinions, inputs, and criticisms, Twitter can be used as a source of data that can be analyzed later [9]. Sentiment analysis is a technique that can analyze individual or group opinions or opinions on a product [10]. Sentiment analysis can be extracted automatically through writing. In this study, users' tweets will determine their emotional state [11]. Sentiment analysis has been used in recent years with the growing popularity of social media, where a company will monitor the reputation of a product and look at the product's brand and public opinion [5]. In the research sentiment analysis can be categorized into positive, negative, and neutral [12].

Research on sentiment analysis of transportation service users has been done previously using the naive bayes analysis method in this study uses TF-IDF and information gain, concluding that the accuracy rate is not reasonable [13]. Furthermore, there are other studies on transportation service users who use The Support Vector Machine (SVM) method with the conclusion that this method is quite suitable for classifying data mining or text mining but can be improved again using advanced machine learning [14]. Furthermore, there is research that Padmapani has carried out in 2018, where the conclusions of this research show that using the Support vector machine method is quite suitable for classifying data mining or text mining but can be improved again using advanced machine learning. This method can improve accuracy by 6% higher than the primary classifier. However, the results are strongly influenced by the main methods' heterogeneity and the dataset's unbalance [7].

Based on the problems and previous research, a system was designed to analyze the sentiment of online transportation services in Indonesia on Twitter social media. The method used is ensemble stacking with TF-IDF feature extraction, which uses base classifier logistic regression and SVM with RBF and linear kernel.
2. RESEARCH METHODOLOGY

2.1. Research Stages

The research was conducted based on a literature study, building a system from problem identification and then testing the system built. The research method in the experiment consists of Literature Study, Problem Identification, System Design, System Implementation, System Test, and Result Analysis. The flow chart in Figure 1. Shows the methodology of the research carried out.

![Figure 1. Research Method](image)

The research method is shown in Figure 1. It is a flow of steps carried out in research, in other words, as a sequence process from the start of the research to the end. In the first process, focus on searching for literature research on sentiment analysis of ensemble stacking and base classifier to be used. After conducting a literature study, the next step is to design an ensemble stacking system for the sentiment analysis. At this stage, we discuss the base classifier, what preprocessing stages are used, and its implementation in the system that will be built later. Next, experiment with the results of the system that has been built, then get the results of the experiments carried out, the system is analyzed and tested on the results obtained based on the results obtained. Parameters that have been determined then draw conclusions based on the experiments that have been carried out.

2.2. System Design

The author designed a system that will be implemented in this study following the analysis of online transportation sentiment analysis data on Twitter social media. Using the ensemble stacking method to improve the analysis results in displaying user sentiment on Twitter media, based on previous research. The following is an overview of the system design that was built:

![Figure 2. Sentiment Analysis Flowchart using the Ensemble Stacking Method](image)
the training and test data will meet again for prediction, cross-validation, and the meta classifier data evaluation. After a series of processes are completed, sentiment analysis results are obtained.

2.3. Term Frequency-inverse Document Frequency

Term Frequency-inverse Document Frequency is the step to convert data from the pre-processing stage into data with values or weights [6]. Where the data are the words that make up the tweet, the purpose of giving this value or weight is to determine how important the word terms are in a sentence [11]. The method used in this research is Term Frequency - Inverse Document Frequency (TF-IDF). This TF-IDF method counts how many words are in a text document. Here is how to calculate the Term Frequency - Inverse Document Frequency (TF-IDF) [3]. At the feature extraction stage, the weight of the tweet values that have been obtained is considered by converting the dataset that has gone through the pre-processing stage into vector data. Moreover, in this study, using the term frequency-inverse document frequency (TF-IDF) weighting extraction method with TF-IDF which calculates the weight of the occurrence of a word in the document [6].

2.4. Ensemble Stacking

The ensemble is the way an algorithm learns data by using a combination of several algorithms. Ensembles are obtained from several classifiers and combine individual results to produce more optimal results [7]. One application of the ensemble algorithm is the ensemble stacking method, where this method uses a different classifier. The first is a base classifier which helps predict from a set of data [6]. The second is the meta classifier, which combines the results of several base classifiers to produce accurate decisions [12]. The following is an illustration of how the ensemble stacking method works:

![Figure 3](image)

**Figure 3.** The Flow chart of the Ensemble Stacking

Based on previous research in sentiment analysis, ensemble stacking increases the basic model's accuracy level [7]. Accuracy will increase in balance with the increase in learning levels, but the execution time will also be longer [2]. In addition, the selection of preprocessing and feature extraction processes also affects sentiment analysis results [12].

2.5. Logistic Regression

Logistic regression is a base classifier method by taking input and multiplying the weight values [12]. Logistic regression can be distinguished by two methods, namely binary logistic regression and multinomial logistic regression [15]. In binary logistic regression, it is a non-linear regression in which there is a model to be determined following the pattern of the curve. Meanwhile, multinomial logistic regression is a regression that is used when the model has a multinomial scale [14]. For example, the following is a logistic regression equation.

\[
\text{Logit}(P) = \ln \left( \frac{P}{1-P} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k
\]

2.6. Support Vector Machines

Support vector machines, commonly abbreviated as SVM, is a machine learning method that takes samples of points in space, where the points in the space are mapped so that samples in other classes can be separated by a more comprehensive gap [15]. In addition to performing linear classification, SVM in performing non-linear classification where this linear classification can be done using the kernel [16]. SMV works when performing non-linear classification by mapping the input that will be processed later into a feature space with high dimensions. Separating the point of a class from another class is obtained by using a hyper lane that has the furthest distance to the closest research data point from any class [17].

3. RESULTS AND DISCUSSION

3.1. Crawling Data

Crawling data is taking information or data from a source, where the source of this research is tweeted on Twitter. This data retrieval requires assistance from the Application Programming Interface (API), but the data obtained is still not structured in CSV format. This study obtained information by entering keywords to get tweets mentioning Gojek and Grab accounts. Figure 3 example of data crawling results.
3.2. Pre-processing

The next step after getting the dataset is pre-processing, which is helpful for cleaning and unstructured processing data. Because the tweet data obtained from the API still uses words that are not standardized, and sometimes there are numbers or symbols. In the pre-processing, case folding, tokenizing, normalization, and word stemming will be carried out so that the data is cleaner and more structured. The following are the stages of the pre-processing process flow that the author did.

![Figure 4. Image Crawling Data](image)

![Figure 5. Preprocessing Flow](image)

<table>
<thead>
<tr>
<th>Stages</th>
<th>Definition</th>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case folding</td>
<td>the stage of changing tweet data from capital letters to lowercase letters and removing all punctuation marks in sentences.</td>
<td>PLS Promo terus ya GOJEK, gue lg ngeributin sm keluarga gue jadi mau gamau harapan gue cm ELO buat anter kmn mnxf0xf9ixa5xvxa @gojekindonesia'</td>
<td>pls promo terus ya gojek, gue lg ngeributin sm keluarga gue jadi mau gamau harapan gue cm elo buat anter kmn mnxf0xf9ixa5xvxa @gojekindonesia'</td>
</tr>
<tr>
<td>URL, Hashtag, Symbol and Removal</td>
<td>This stage removes symbols such as hashtags from the dataset.</td>
<td>[pls, 'promo', 'terus', 'ya', 'gojek', 'gue', 'lg', 'ngeributin ', 'sm', 'keluarga', 'gue', 'jadi', 'mau', 'gamau', 'harapan', 'gue', 'cm', 'clo', 'buat', 'anter', 'kmn', 'mn']</td>
<td>[pls, 'promo', 'terus', 'ya', 'gojek', 'gue', 'lg', 'ngeributin ', 'sm', 'keluarga', 'gue', 'jadi', 'mau', 'gamau', 'harapan', 'gue', 'cm', 'clo', 'buat', 'anter', 'kmn', 'mn']</td>
</tr>
<tr>
<td>Tokenizing</td>
<td>Each word will be separated based on the space found in a process on tweet data.</td>
<td>[tolong, 'promo', 'gojek', 'ngeributin ', 'keluarga', 'gamau', 'harapan', 'Cuman', 'kamu', 'anter', 'kemana']</td>
<td>[tolong, 'promo', 'gojek', 'ngeributin ', 'keluarga', 'gamau', 'harapan', 'Cuman', 'kamu', 'anter', 'kemana']</td>
</tr>
<tr>
<td>Normalisasi</td>
<td>correcting inappropriate words and abbreviations that are not commonly used in everyday life. For that, the words are corrected first.</td>
<td>['tolong', 'promo', 'gojek', 'ngeributin ', 'keluarga', 'gamau', 'harapan', 'Cuman', 'kamu', 'anter', 'kemana']</td>
<td>['tolong', 'promo', 'gojek', 'ngeributin ', 'keluarga', 'gamau', 'harapan', 'Cuman', 'kamu', 'anter', 'kemana']</td>
</tr>
</tbody>
</table>

3.3. Data labeling

Data labeling is done to explain whether tweets from Twitter users are included in the negative or positive category. The author only categorizes the labels from the dataset into two categories, positive and negative. In this labeling process, the writer does it manually with the help of two other people for the labeling process where voting will be carried out in deciding to make tweets from this user, whether positive or negative, later. The results of the
3.4. Split Data

Data split is a method of dividing the dataset into two parts: train data and test data. The data split method or data splitting itself is used to evaluate the learning methods' results. Where the splitting of the dataset is carried out in certain propositions. The following is an explanation of the train data and test data:

A. Data Training

Information data used to train an algorithm training. The training data includes input and output data that are expected to match. The training data is separate so that it can be used to train and test the model that will be used to analyze online transportation sentiment later.

B. Data Testing

The test data is an approximate reference in seeing how well the results obtained from the train data are in the model's performance. In this case, it is seen from the two data, namely train data and test data, which have the same label value.

3.5. Feature Extraction

At the feature extraction stage, the weight of the tweet value that has been obtained is considered by converting a dataset that has gone through the pre-processing stage to become vector data. Furthermore, in this final project, using the term frequency-inverse document frequency (TF-IDF) weighting extraction method with TF-IDF calculating the weight of the occurrence of a word in the document. Here's how to calculate Term Frequency - Inverse Document Frequency (TF-IDF):

\[ W_{df} = tf_{df} \cdot \log\left( \frac{n}{df} \right) \]  

Where \( w \) in weight of calculation, \( df \) is word of a document, \( D \) is The total number of documents to be calculated, \( df = \) number of words in the document (D), \( t = \) word that has been marked in (df).

3.6. Classification

A. Base Classifier

In the base classifier, two basic classifiers will be applied, logistic regression and SVM. In SVM, the author uses the RBF kernel, which is also linear in the basic SVM classifier. Assuming the use of these two main classifications, because many sentiment analysis studies use these two classifications, it provides performance and diversity compared to other classifiers. With better performance and many studies that have used two main classifications (SVM and logistic regression), it produces even better performance when used in ensemble stacking.

B. Meta Classifier

The meta classifier that will be built in this study uses logistic regression. In this meta classifier stage, the trained input data is data from the previous base classifier. The meta classifier will predict the results of the data from the previous two base models and will return a final prediction depending on the three predictions of the base classifier. The sci-kit-learn library makes implementing standard classifiers in the Python programming language accessible. The author uses the SVM algorithm with the RBF kernel in the meta classifier used using the SVM algorithm.

C. Evaluation

Evaluations are used to see the suitability of the model with the data. A confusion matrix will measure the performance of the model that has been built in the form of a table. With this table, accuracy, precision, recall, and f1-score can be observed as a benchmark for the system's performance. In addition, the confusion matrix table will make it easier to see the test data that has been classified, where the author can see the number of correct and incorrect tests in the classification that has been carried out.

3.7. Test Scenario 1

In the first scenario, testing of data splitting is carried out. This test aims to see the accuracy results using the best ensemble stacking method in the distribution of train data and testing data. In the test, the split data will be divided into two different ratios: (a) In this scenario, data with a ratio of 80:20 is used. (b) This scenario uses data with a ratio of 70:30. This first scenario uses the SVM meta classifier with a linear kernel. The test results in scenario one can be seen in Tables 2 and 3.
In the first scenario, retrieval of the grab dataset uses ensemble stacking with a linear kernel on the gojek dataset gets an accuracy of 85%. However, in the second scenario, it can be seen in the grab dataset where the results are different. SVM with linear kernel got 85% accuracy, SVM with RBF kernel got 86% accuracy, and logistic regression achieved 85% accuracy. Ensemble stacking affects the level of accuracy to be better than the base classifier used, with an average increase in the level of accuracy of 3% obtained from the system that has been built. In the first scenario, retrieval of the grab dataset uses ensemble stacking with data splitting at a ratio of 80:20 to get 86% accuracy. Compared to using an SVM-based classifier with an RBF kernel to get 83% results, SVM linear kernel gets 84% accuracy. In logistic regression, the result was 83%. The results obtained in the ensemble stacking method get an increase of about 3% from the existing base classifier. And in the first scenario, by comparing the results obtained on the data split with a ratio of 70:30, the Gojek dataset gets 87% results, and the grab dataset gets 85% accuracy from the data splitting results, getting a decrease in results from the 80:20 ratio, the actual level of accuracy is 1% of each data set. What makes 80:20 better than the 70:30 accuracy rate is a more significant proportion of the training data that allows more information and diversity to be derived from the data set. A more significant proportion of the training data provides more information that can be obtained.

For the second test scenario, by changing the meta classifier to logistic regression, previously using SVM with a linear kernel. The accuracy results obtained in the stacking ensemble on the gojek dataset is 88%, with a data splitting ratio of 80:20. Moreover, the results are 87% accuracy with 70:30 data separation. The results in the second scenario are not visible when using the logistic regression meta classifier on the gojek dataset. The base classifier gets the same results as logistic regression, and SVM with a linear kernel on the gojek dataset gets an accuracy of 85%. However, in the second scenario, it can be seen in the grab dataset where the results are different from the basic logistic regression classifier with the SVM linear kernel. This is because the linear SVM kernel gets better accuracy results than logistic regression. The result of ensemble stacking is also influenced by the meta classifier selector used by selecting the next meta classifier.

### 3.8. Test Scenario 2

In the third scenario, testing is carried out by comparing the different meta-classifiers used. The algorithm used in the meta classifier used is SVM with linear kernel and logistic regression. The selection of the two algorithms is based on the results obtained by the algorithm in the previous base classifier. The test results can be seen in Tables 8 and 9.

### 3.9. Analysis of Test Results

For the first test scenario, the ensemble stacking method is used by testing the splitting of the gojek dataset with a ratio of 80:20 and 70:30. The results obtained are 88% at a ratio of 80:20 by comparing the effect on the base classifier that has been done previously to get better results. SVM with linear kernel got 85% accuracy, SVM with RBF kernel got 86% accuracy, and logistic regression achieved 85% accuracy. Ensemble stacking affects the level of accuracy to be better than the base classifier used, with an average increase in the level of accuracy of 3% obtained from the system that has been built. In the first scenario, retrieval of the grab dataset uses ensemble stacking with data splitting at a ratio of 80:20 to get 86% accuracy. Compared to using an SVM-based classifier with an RBF kernel to get 83% results, SVM linear kernel gets 84% accuracy. In logistic regression, the result was 83%. The results obtained in the ensemble stacking method get an increase of about 3% from the existing base classifier. And in the first scenario, by comparing the results obtained on the data split with a ratio of 70:30, the Gojek dataset gets 87% results, and the grab dataset gets 85% accuracy from the data splitting results, getting a decrease in results from the 80:20 ratio, the actual level of accuracy is 1% of each data set. What makes 80:20 better than the 70:30 accuracy rate is a more significant proportion of the training data that allows more information and diversity to be derived from the data set. A more significant proportion of the training data provides more information that can be obtained.

For the second test scenario, by changing the meta classifier to logistic regression, previously using SVM with a linear kernel. The accuracy results obtained in the stacking ensemble on the gojek dataset is 88%, with a data splitting ratio of 80:20. Moreover, the results are 87% accuracy with 70:30 data separation. The results in the second scenario are not visible when using the logistic regression meta classifier on the gojek dataset. The base classifier gets the same results as logistic regression, and SVM with a linear kernel on the gojek dataset gets an accuracy of 85%. However, in the second scenario, it can be seen in the grab dataset where the results are different from the basic logistic regression classifier with the SVM linear kernel. This is because the linear SVM kernel gets better accuracy results than logistic regression. The result of ensemble stacking is also influenced by the meta classifier selector used by selecting the next meta classifier.

### Table 2. Test Results of Gojek Meta Classifier Dataset with Linear Kernel SVM Meta Classifier

<table>
<thead>
<tr>
<th>Scenario 1</th>
<th>Accuracy</th>
<th>Recall</th>
<th>Precision</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rasio 80:20</td>
<td>88%</td>
<td>86%</td>
<td>90%</td>
<td>87%</td>
</tr>
<tr>
<td>Rasio 70:30</td>
<td>87%</td>
<td>85%</td>
<td>88%</td>
<td>86%</td>
</tr>
</tbody>
</table>

### Table 3. Test Results of Grab Meta Classifier Dataset with Linear Kernel SVM Meta Classifier

<table>
<thead>
<tr>
<th>Scenario 1</th>
<th>Accuracy</th>
<th>Recall</th>
<th>Precision</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rasio 80:20</td>
<td>86%</td>
<td>85%</td>
<td>85%</td>
<td>85%</td>
</tr>
<tr>
<td>Rasio 70:30</td>
<td>85%</td>
<td>84%</td>
<td>84%</td>
<td>84%</td>
</tr>
</tbody>
</table>

### Table 4. Gojek Dataset Test Results with Meta Classifier Logistic Regression

<table>
<thead>
<tr>
<th>Scenario 2</th>
<th>Accuracy</th>
<th>Recall</th>
<th>Precision</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rasio 80:20</td>
<td>88%</td>
<td>86%</td>
<td>89%</td>
<td>87%</td>
</tr>
<tr>
<td>Rasio 70:30</td>
<td>87%</td>
<td>85%</td>
<td>88%</td>
<td>86%</td>
</tr>
</tbody>
</table>

### Table 5. Grab Dataset Test Results with Meta Classifier Logistic Regression

<table>
<thead>
<tr>
<th>Scenario 2</th>
<th>Accuracy</th>
<th>Recall</th>
<th>Precision</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rasio 80:20</td>
<td>85%</td>
<td>84%</td>
<td>84%</td>
<td>84%</td>
</tr>
<tr>
<td>Rasio 70:30</td>
<td>84%</td>
<td>84%</td>
<td>83%</td>
<td>84%</td>
</tr>
</tbody>
</table>

### 4. CONCLUSION

Based on research conducted with two test scenarios, it can be said that the application of the ensemble stacking method to online transportation sentiment analysis on Twitter, data distribution training, and data testing of ensemble system accuracy and selection of meta classifiers are also essential to get better accuracy from compiling ensembles. From the results that have been carried out on sentiment analysis research on Twitter on online transportation in Indonesia using ensemble stacking, the best results obtained on the gojek dataset are 88% accuracy in the SVM kernel linear meta classifier and logistic regression with training data distribution and 80:20
test data. The best result on the fetch dataset is the 86% accuracy rate obtained when using the linear SVM meta classifier with training data distribution and test data of 80:20. The thing that affects the difference in the accuracy of the results of the ensemble stacking system from the gojek and grab datasets is the heterogeneity of the gojek and grab datasets. In addition to the heterogeneous dataset, the preprocessing process will also affect the results. The distribution of the dataset can be seen from the results of the distribution of the 80:20 training data distribution, which is always better than the 70:30 distribution, where the more training data that is classified, the better the accuracy obtained.

REFERENCES


