Sentiment analysis of GO-JEK services quality using Multi-Label Classification

A Amalia*1, D Gunawan2, K Nasution3

1Department of Computer Science, Universitas Sumatera Utara, Jl. dr. Mansur No. 9 Kampus USU Medan 20155

2Department of Information Technology, Universitas Sumatera Utara, Jl. dr. Mansur No. 9 Kampus USU Medan 20155

3Department of Informatic, Universitas Islam Sumatera Utara, Jl. SM-Raja – Teladan Medan

*E-mail: amalia@usu.ac.id

Abstract. One of the most popular online transportation providers in Indonesia is GO-JEK. At the first establishment, GO-JEK is only an online taxi motorbike service (in Bahasa Indonesia: ojek) that transforms from a conventional taxi motorbike. After several years GO-JEK began to develop more services, like GO-FOOD, GO-SEND, GO-CAR, GO-MART, GO-RIDE, GO-PAY, GO-TIX, GO-BOX, and GO-MED. As the GO-JEK services develop into more categories, it is more challenging to automatically analyze the sentiment polarity for each category of services category. An ordinary classification algorithm or single-label classification is concerned with learning from a set of examples associated with a single label classification. However, in this case, we want to classify GO-JEK services based on two-class targets, which are GO-JEK service categories and polarity sentiment classification. This research methodology contains Dataset Preparation, Feature Selection, Basic text mining process, train, and split dataset, and Classification. We implemented two classification methods: Multi-Label Classification and a simultaneous classification using the Random Forest Algorithm as a comparison. Based on this dataset, the most mentioned GO-JEK service is GO-FOOD followed by GO-SEND, GO-RIDE, GO-CAR, and GO-MART. Based on the service category, GO-FOOD gets the most positive reviews following by GO-RIDE and GOJEK 90. Some service categories like GO-SEND and GO-MART get more negative reviews than positive reviews. The accuracy of the Multi-label classification method raised 76%. Simultaneous classification accuracy using the Random Forest algorithm produce for service category classification yields 97% and only 78% for sentiment polarity classification. We can see that both algorithm, multi-label classification, and Random Forest algorithm, yields almost the same classification accuracy for polarity sentiment classification. We can conclude that the imperfect accuracy in polarity sentiment classification is related to the difficulty of identifying the most suitable polarity for each tweet. Many sarcasm tweets and slang words that cannot be identified.

1. Introduction
The development of Indonesian transportation has entered a disruptive model. Online transportation, such as Taxi Online, becomes more popular because customers can get many conveniences and ease in one click away from their smartphone. Furthermore, users can search for the most suitable transportation services in their need based on reviews of other customers from the application or social media. Many people commonly use social media to share their opinion and comments. Many companies utilized these reviews as feedback to increase their services and the quality of their products. One of the most popular online transportation providers in Indonesia is GO-JEK. At the first establishment, GO-JEK is only an online taxi motorbike service (in Bahasa Indonesia: ojek) that transforms from a conventional taxi motorbike. After several years GO-JEK began to develop more services, like GO-FOOD (buying foods services), GO-SEND (shipping goods in the city), GO-CAR (car taxi online), GO-MART (shopping services), GO-RIDE (motorbike taxi ), GO-PAY (virtual money that can be used to pay GO-JEK services transaction), GO-TIX (buying cinema movie services), GO-BOX (moving house services) and GO-MED (buying medicine services) [1]. GO-JEK has an official Twitter account, namely @gojekindonesia. The purpose of this account is to accommodate the reviews and comments from their customers, particularly about GO-JEK services. Based on these reviews, companies can analyze a customer's opinion to improve their products.
and services. Opinion analysis or sentiment analysis is contextual mining of text to identify and classify opinions or sentiment in the source text, whether underlying positive, neutral or negative. Many previous studies analyze the sentiment with machine learning approaches. As the GO-JEK services develop into more categories, it is more challenging to analyze each service category's sentiment polarity. An ordinary classification algorithm or single-label classification like Support Vector Machine (SVM), Naïve Bayes, and Random Forest Classifier are concerned with learning from a set of examples associated with a single label classification. However, in this case, we want to classify GO-JEK services based two-class target, which is GO-JEK service categories and polarity sentiment classification for each GOJEK services categories. One solution we can implement a single label classification algorithm with two simultaneous steps. We classify the service categories and follow the next step, a classification based on the polarity sentiment. Another solution we can implement a multi-label classification algorithm. Multi-label classification is an algorithm to classify a set of examples related to more than one class. The aim is to predict multiple output variables for each input instance. This differs from the binary or multi-class classification, which involves only a single target variable [2]. This study aims to classify GO-JEK reviews into its service category and analyze them into positives, negatives, and neutral polarity for each category of the services. We implemented a multi-label or multi-output classification algorithm using the Scikit-learn classifier platform [3][4]. Scikit-learn is an integrated module for Python, especially for machine learning algorithms [3]. In this study, we also implemented simultaneous steps classification of a single label classification algorithm, which is a Random Forest algorithm as a comparison. The paper is organized as follows. Related works are given in Section 2. The methodology is described in Section 3. Section 4 presents the Result and Discussion. The Conclusions are presented in Section 5.

2. Related Work

Some of the previous studies implemented multi-label classification for sentiment analysis, like [5][6][7]. The study by [5] has three main components, which are text segmentation, feature extraction, and multi-label classification. The study by [5] utilized three different dictionaries and implemented a bag of words for feature representation. The study by [6] proposed a methodology for expanding a multi-label emotion lexicon based on a collection of un-labeled tweets using multi-label classification. The study by [7] proposed the emotion concentration to reduce the issue of noisy training documents and identify the same word's emotional senses with exploited topic models. Our work classifies the dataset only has one polarity sentiment among three classes, positive, neutral or negative. These studies observed that the dataset could have another sentiment class like "happy" or "sad." Another differentiation with our study is the observed language. Our study observed Bahasa Indonesia that have a different pre-processing task with English, especially in the stemming and normalization stage. Some previous studies observed Bahasa Indonesia like the study by [8][9].

3. Methodology

The methodology for mining and classify the reviews about GO-JEK services quality contains several processes. The first process is dataset preparation and following by another process. More detail about this methodology can be seen in general architecture in Figure 1.
3.1 Dataset Preparation

There are two subprocesses in preparing the dataset stage: the Crawling Data and Class Labeling process. Crawling Data is a process to collect the data. We collected the user's review of several GO-JEK categories from Twitter. These reviews or the post on Twitter also known by "tweet". We choose Twitter because Indonesia, as one of the countries that have the most users using Twitter. Moreover, Twitter as one of the social media that provide Application Programming Interface (API) that make researchers easy to crawl the data. We crawled or collected about 1500 review tweets that mention @Gojekindonesia. The next step is the Class Labeling process. Class Labeling is a process to label the tweets into two suitable classes or categories. The first label is to classify the tweets based on the GO-JEK service category. This study only took six GO-JEK service categories: GO-CAR, GO-SEND, GO-MART, GO-FOOD, GO-RIDE, and GO-PAY. We also classify the tweets that do not mention specific GO-JEK services but contain sentiment polarity to general GO-JEK services into class GOJEK. In this process, we implemented a string matching method to do this labeling process automatically. The label of these service categories is based on the keyword that mentions the tweets' service category. We also consider the possibilities of the typo from the customers to mention these categories. For example, "gocar", "go-car" and "go car" mean the same, that the users are talking about Go-Car Services. The other labeled process is to classify Twitter based on their sentiment polarity, positive, neutral, and negative. In this process, we utilized a human annotator and did this process manually.

3.2 Feature Selection

Feature Selection is a process to prepare the dataset, so machine learning approaches can further process it. Generally, machine learning can only process the data as input in number or vector form. Therefore, we need to pre-process the dataset from the previous stage that is still in text form. Several stages in this process are Data Cleaning, Case Folding, Tokenization, and Stop Word Removal. Data Cleaning is a process to remove noise or unwanted terms or symbols from the dataset. In this process, we remove all the symbols and numbers like "@", "RT", "#" and emoticon. We also remove URL addresses and punctuations. The next step is Case Folding, or converting all the text in the dataset into lower case. The next process is Tokenization; this process is a process to split the Twitter sentences into words or terms. Stop Word Removal process is a process to remove all the vocabularies in the dataset that include in stop-words vocabularies. Stop-words vocabularies are words that are claimed as not important words or does not have much effect in a language. In this study, we implemented Tala Vocabularies [10] as stop-words vocabularies for Bahasa Indonesia.

3.3 Convert Negation

Convert negation is a process to identify all negative terms that appear in front of a word. The purpose of this identification is to convert the polarity of the word into its opposite. Because all the negative terms like "bukan", "tidak", "enggak", "ga", "jangan", "nggak", "tak", and "gak" (in English: "Not") that appears before a word will change the polarity value of the word, in this stage, we convert this sequence word into the antonym of the adjective word. For example, the sequence word "tidak bagus" (in English: not good), we will replace "tidak bagus" into "jelek" (in English: bad). An example of this process can be seen in Table 1.

| Before Convert Negation | After Convert Negation |
|-------------------------|------------------------|
| Tidak bagus [negative] x [positive] | Jelek [negative] |

3.4 Stemming

Stemming is a process to get a root base of a word by removing all the words' affixes. This process aims to minimize the mistake in identifying the keywords from the general word vocabularies. We implemented Nazief Adriani algorithm [11], which is one of the best stemming algorithms for Bahasa Indonesia. We utilized a dictionary to check general words in Bahasa Indonesia; this dictionary contains 29932 words. We get this dictionary by downloading it from Bahasa Indonesia dictionary on kateglo.com [12].

3.5 Normalization

Normalization identifies informal words in the dataset and converts these words into standard language form. Usually, informal words are called as slang words. Many slang words appeared on an informal platform like social media. Moreover, one of the challenges of analyzing social media is too many slang words in it. We can not analyze this kind of words because they are out of the vocabulary and have a higher tendency in ambiguity.
interpretation. For example, the word "betul" (in English: True), sometimes in slang words will be written as "btl", "betol", "bnr" "tul" etc. This problem can be solved by replacing the informal words with a formal word with the same context. Besides normalizing the slang words, in this study, we also normalize the acronym and abbreviation. The excerpt of this process can be seen in Table 2.

### Table 2. The Excerpt of Synonym Dictionary

| Informal Words | Synonym in Formal Words |
|----------------|-------------------------|
| Bener          | Benar                   |
| Blom           | Belum                   |
| Sampe          | Sampai                 |
| Dapet          | Dapat                 |
| Btl            | Betul                  |

#### 3.6 Data Representation

Data representation is a process to convert text data from feature selection into numbers or vectors. This process is also known as the word weighting process. In this study, we implemented a term frequency method. All terms or words are represented based on how many times they appear on a tweet.

#### 3.7 Train and Split Dataset

Train and split test a utility to create training and testing sets from a single data set randomly. We implemented utility from Scikit Learn for Python. This study split the dataset 70% as a training set and 30% tweets as a testing set.

#### 3.8 Classification

In this stage, we implemented Multi-Label classification and Random Forest Classification as a comparison. We implemented MultiOutputClassifier Algorithm from Scikit Learn that predicts multiple variables of output from data input. This algorithm is one of the Multioutput-multiclass classification and multi-task classification algorithms modules in Scikit Learn. Multitask-multiclass classification and multi-task classification means that a single estimator has to handle several common classification tasks. This is both a generalization of the multi-label classification task, which only considers binary classification, and a generalization of the multi-class classification task [2]. This method is also known as multi-label learning, multi-dimensional, or multi-objective [13]. Multi-output models can produce multiple outputs to assign to multiple target variables y for each test instance [2]. For instance, an example of fruit classification, a sample data could be labeled as "pear"; and "yellow" or "green" for another output. This example is similar to our study case, which classifies the GO-JEK service categories and sentiment polarity for each category. This is the simultaneous prediction of multiple target variables using the same set of predictive variables. The prediction task is called a multi-label classification because the target variables are binary and not real-valued variables [14].

As a comparison, we also implemented a simultaneous process classification using Random Forest Classifier. We choose Random Forest because this algorithm easy to use, and many previous studies showed this algorithm generates a great result. First, the system will predict the GOJEK service category, and then the system will predict the polarity sentiment for each category.

#### 4. Result and Discussion

In this study, the Preparing Dataset stage is a process to yield a collection of tweets with two label classification, which is based on GO-JEK service category and based on sentiment polarity category. The crawling process yields a tweets dataset that contains 370 positive reviews, 370 neutral reviews, 560 negative reviews, and 200 unrelated tweets to GO-JEK. To make the data balance; therefore, as many as 190 negatives, reviews were randomly excluded from the dataset. We also exclude 200 unrelated tweets because not related to Go-JEK quality services reviews. After this process, the number of twitters in this dataset is 1110 tweets. Based on this dataset, the most mentioned GO-JEK service is GO-FOOD, which are 267 tweets, followed by GO-SEND are 266 tweets, GO-RIDE is 236 tweets, GO-CAR is 202 tweets, and GO-MART are 20 tweets. The percentages of the distribution of these service categories can be seen in Figure 2.
Based on the service category, GO-FOOD gets the most positive reviews, which are 125 tweets following by GO-RIDE 100 tweets and GOJEK 90 tweets. Some service categories like GO-SEND and GO-MART get more negative reviews than positive reviews. The distribution of sentiment polarity for each service category can be seen in Figure 3.

Feature selection is a process to extract the most important feature from the dataset. In this study, we implemented a Bag of Words (BoW) feature method. We split all the twitters from the dataset into words by words. We implemented some basic text mining like case folding, stemming, normalization, and stop word removal from reducing the dimension. The excerpt of this stage can be seen in Table 3.

| AMOUNT | SERVICE CATEGORY |
|--------|------------------|
| 125    | GO-FOOD          |
| 100    | GO-RIDE          |
| 90     | GOJEK            |
| 32     | GO-SEND          |
| 16     | GO-CAR           |
| 14     | GO-PAY           |
| 11     | GO-MART          |

**Figure 2.** The percentages of the distribution GO-JEK service categories

![Pie chart showing the distribution of GO-JEK service categories.]

**Figure 3.** The Distribution of Sentiment Polarity for Each Service Category

![Bar chart showing the distribution of sentiment polarity for each service category.]

**Table 3.** The Excerpt of Feature Selection

| tweet | polarity | category | hasil preprocessing |
|-------|----------|----------|--------------------|
| @gojekindonesia saya Di bontang dan dikasih vo... | negatif | GO-CAR | [‘bontang’, ‘kasih’, ‘voucher’, ‘gocar’, ‘lan’] |
| @gojekindonesia Kamarin saya dibohongin driver... | negatif | GO-CAR | [‘kamarin’, ‘dibohongin’, ‘driver’, ‘gocar’, ‘lan’] |
| Telat ngurusin hal penting karna driver go-car... | negatif | GO-CAR | [‘telat’, ‘ngurusin’, ‘penting’, ‘karna’, ‘di’] |
| @gojekindonesia Kadang DRIVER menas dirugikan... | negatif | GO-CAR | [‘kadang’, ‘driver’, ‘mas’, ‘nugi’, ‘bilas’, ‘lan’] |
| @yudobat @gojekindonesia Kejinya emg lagi error... | negatif | GO-CAR | [‘yudobat’, ‘kek’, ‘emg’, ‘error’, ‘mas’, ‘cob’] |
We split the dataset into 70% training and 30% for the testing stage. The accuracy of the Multi-label classification method is 76% for testing accuracy. Simultaneous classification accuracy using the Random Forest algorithm produces for service category classification yields 94% for testing accuracy. For polarity, sentiment classification yields 78% for testing accuracy. We can see that both algorithms, multi-label classification, and Random Forest algorithm yields almost the same classification accuracy. The confusion matrix of Random Forest accuracy can be seen in Figure 4 and Figure 5. Based on Figure 5, we can see that almost all category classification has high accuracy.

Figure 4. Confusion Matrix of GOJEK Service Categories Classification

Figure 5. Confusion Matrix of Polarity Sentiment Classification
Figure 5 shows the confusion matrix of polarity sentiment classification. We can see that some of the predicted classifications are different from actual values. The best prediction is for negative polarity sentiment classification, which has correct 74 predictions of 92 data testing. The worst prediction is the Neutral prediction class, which has only 78 correct 103 data testing predictions. The poor accuracy in polarity sentiment classification is related to the difficulty of identifying the most suitable polarity for each tweet—many sarcasm tweets and slang words that can not be identified.

5. Conclusion
This study aims to analyze the reviews about GO-JEK services on Twitter and classify them into positives, negatives, and neutral polarity for each category of the services. We implemented a multi-label or multi-output classification algorithm using the Scikit-learn classifier platform. In this study, we also implemented simultaneous steps classification of a single label classification algorithm, which is the Random Forest algorithm as a comparison. The preparing Dataset stage is a process to yield a collection of tweets with two label classification, which is the GO-JEK service category and polarity sentiment category. Based on the dataset, the most mentioned GO-JEK service is GO-FOOD, followed by GO-SEND, GO-RIDE, GO-CAR, and GO-MART. Based on the service category, GO-FOOD gets the most positive reviews following by GO-RIDE and GOJEK. Some service categories like GO-SEND and GO-MART get more negative reviews than positive reviews. The accuracy for Multi-label classification is 76% for testing accuracy. Simultaneous classification accuracy using the Random Forest algorithm produces for service category classification yields 94% for testing accuracy. For polarity, sentiment classification yields 78% for testing accuracy. We can see that both algorithms, multi-label classification, and Random Forest algorithm yields almost the same classification accuracy. The reduced accuracy in polarity sentiment classification is related to the difficulty of identifying the most suitable polarity for each tweet. Many sarcasm tweets and slang words that can not be identified.

References
[1] Anon Tentang Kami | GO-JEK Indonesia
[2] Montiel J, Read J, Bifet A and Kegel B 2018 Scikit-Multiflow: A Multi-output Streaming Framework J. Mach. Learn. Res. 19 1–5
[3] Pedregosa Fabian, Varoquaux Gaël, Gramfort Alexandre, Duchesnay Edouard, Pedregosa Fabian, Grisel Olivier, Blondel Mathias, Prettenhofer Pierre, Weiss R, Vanderplas J, Cournapeau D, Brucher Matthieu, Perrot Edouard and Dubourg Vincent 2011 Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos Fabian Pedregosa, Gaël Varoquaux, Alexandre Passos, Edouard Duchesnay, Matthieu Perrot J. Mach. Learn. Res. 12 2825–30
[4] Anon 1.12. Multiclass and multilabel algorithms — scikit-learn 0.20.2 documentation
[5] Liu S M and Chen J-H 2015 A multi-label classification based approach for sentiment classification Expert Syst. Appl. 42 1083–93
[6] Bravo-Marquez F, Frank E, Mohammad S M and Pfahringer B 2016 Determining Word-Emotion Associations from Tweets by Multi-label Classification 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI) (IEEE) pp 536–9
[7] Xin Li, Haoran Xie, Yanghui Rao, Yanjia Chen, Xuebo Liu, Huan Huang and Fu Lee Wang 2016 Weighted multi-label classification model for sentiment analysis of online news 2016 International Conference on Big Data and Smart Computing (BigComp) (IEEE) pp 215–22
[8] Amalia A, Oktinas W, Aulia I and Rahmat R F 2018 Determination of quality television programmes based on sentiment analysis on Twitter 2nd International Conference on Computing and Applied Informatics 2017 (IOP Publishing)
[9] Amalia A, Lydia M S, Fadilla S D, Huda M and Gunawan D Document Clustering Optimization with Synonymy Dictionary Check Function Case Study: Documents in Bahasa Indonesia
[10] Tala F Z A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia
[11] Adriani M, Asien J, Nazief B and Williams H E 2007 Stemming Indonesian: A Confident-x-Stripping Approach ACM Trans. Asian Lang. Inf. Process. 6 1–33
[12] Anon Kateglo ~ Kamus, tesaurus, dan glosarium bahasa Indonesia
[13] Read J, Reutemann P, Pfahringer B and Holmes G 2016 Meka: A Multi-label/Multi-target Extension to Weka J. Mach. Learn. Res. 17 1–5
[14] Spyromitros-Xioufis E, Tsoumakas G, Groves W and Vlahavas I 2012 Multi-Label Classification Methods for Multi-Target Regression arXiv Prepr.