# Twitter Sentiment Analysis of Online Transportation Service Providers

Sonia Anastasia and Indra Budi
Faculty of Computer Science
Universitas Indonesia
Depok, Indonesia
Email: sonia.anastasia@ui.ac.id and indra@cs.ui.ac.id

***Abstract*—GO-JEK and Grab are two most popular online transportation service providers in Indonesia. The competition between the two is tight to acquire new customers and provide the best service. No official survey has been conducted in measuring customer satisfaction at both companies to give insight which company offers better service quality. This study aims to measure GO-JEK and Grab customer satisfaction through sentiment analysis of Twitter's data. Both companies use Twitter to reach their customers and promote their service. We collect 126,405 tweets from February to March 2016 containing GO-JEK and Grab keywords. Then, we pre-process the tweets and label manually before they are classified using three algorithms: Support Vector Machine, Naïve Bayes, and Decision Tree. We made and compared several classification schemes with different keywords, dataset, and *k-fold* cross validation techniques. Finally, we calculate Net Sentiment Score which correlates with customer satisfaction using classification results. The experiments shows that Grab's customer satisfaction is higher than GO-JEK's. The study also shows that customers tend to mention both companies Twitter account for bad experiences and not mentioning company's account for positive comments.**

***Keywords—sentiment analysis, GO-JEK, Grab, Support Vector Machine, Naïve Bayes, Decision Tree, Net Sentiment Score.***

## I. INTRODUCTION

Smartphone technology develops rapidly in the past few years. This is being followed by increasing popularity of social media and online transportation service application. Social media has become more popular and is widely used by people of all ages. Twitter is one of the most popular social media with 310 million active users monthly[1]. Many organization, business companies specifically, use this social media in order to gain some benefit for their business.

Social media marketing is a new way to promote products and increase revenue and reach more customers with less budget and resources. Social media marketing gives benefit for business by reducing cost and staff time, and increasing revenue generation [1]. According to Social Media Marketing Industry Report 2015, the benefits of Social Media Marketing are increased exposure to business, increased traffic, developed loyal fans, provided marketplace insights, and generated leads [2]. Some studies also show that companies can measure customer satisfaction and customer insight by analyzing social media data [3,4].

This phenomenon also happens in Indonesia, many companies use social media to support their business activity. Two leading online transportation service provider, GO-JEK and Grab use twitter for supporting their business and one of customer service tool. Both companies have apps available in Android and iOS and can be downloaded for free. GO-JEK and Grab have been downloaded approximately 5 million times[2] and has Twitter account at @gojekindonesia and @GrabID respectively. 5 million downloads infers that both apps have a lot of users and both are main choices of transportation for people.

The purpose of this study is to measure customer satisfaction of GO-JEK and Grab, by analyzing their Twitter data. We analyze Twitter data by conducting sentiment analysis to tweets containing keywords about GO-JEK and Grab. Each tweet is classified by three algorithms with several classification schemes. The experiment with highest classifier performance then is used to calculate Net Sentiment Score (NSS). NSS is a method to measure customer satisfaction from social media data. NSS score can be any number between -100 and 100, higher score indicates better customer satisfaction. We use NSS to indicate the customer satisfaction based on a study by Stelzner that NSS shows same result as traditional customer satisfaction measurement method [5].

The result of this study can benefit GO-JEK and Grab to measure their customer's satisfaction. GO-JEK and Grab get feedback from their customers thus they can improve their service quality to satisfy customers.

The rest of this paper as follows. We explain the related work in next section and method that have been used in

[1] https://about.twitter.com

[2] https://play.google.com/

 978-1-5090-4629-4/16/$31.00 

section 3. The results and analysis is discussed in section 4 and finally section 5 conclude the finding of this research.

## II. Related Works

Sentiment analysis of social media has been widely studied. Vidya et.al measured Net Brand Reputation (NBR) of three mobile service providers in Indonesia using Support Vector Machine, Naïve Bayes, and Decision Tree algorithms [3]. They compared the score of NBR which use social media data, and Net Promotor Score (traditional approach) to measure customer satisfaction. The result shows that NBR does not correlate with NSS.

Zhang et.al analyzed comments written in Mandarin on an application in Appstore [4]. They grouped the comments based on its rating and length. They combined Naïve Bayes Multinomial and Support Vector Machine LibLinear algorithms with N-gram feature to classify data. One of the finding is longer comments are more difficult to be classified into a sentiment category.

Chamlertwat et.al proposed Micro-Blog Sentiment Analysis System (MSAS), a combination of machine-learning and lexicon-based approach to analyze smartphone customer's behavior in social media [6]. Support Vector Machine as machine-learning algorithm classifies data to Opinion and Non-Opinion category. Then, they calculated positive and negative polarity score of Opinion data based on SentiWordNet 3.0 lexicon. They also extracted feature in each sentence to get richer information of a product. MSAS is validated by experts in smartphone industry.

Yi Zhang and Peter Desousa compared the performance of five algorithms; Sscore Classifier, Support Vector Machine, Maximum Entropy, Weighted Support Vector Machine, and Weighted Maximum Entropy, to classify data from different sources with different characteristic; Twitter, Amazon, and Movie Reviews [7]. Classification result was compared with human annotator's classification. The result shows that different algorithms are working best for different data sources.

All previous studies discussed above analyzed social media data with machine-learning algorithm and supervised-learning approach. Our study has similarities with [3] in algorithms used and aims of study to analyze customer satisfaction. SVM, Naïve Bayes, and Decision Tree worked well in [3] and since our study has the same purpose, we use the same algorithms. [4], [6], [7] also used SVM and the performance is acceptable. Naïve Bayes is used in [4] and the performance is better than SVM. Decision Tree was only implemented in [3] and the performance is good. We consider classification result in [3], [4], [6], and [7] in selecting algorithms for this study. We analyze Twitter data sentiment to measure customer satisfaction of GO-JEK and Grab. We test three classification algorithms; Support Vector Machine, Naïve Bayes, and Decision Tree. We create several classification schemes combining different kind of keywords and dataset, and K-fold cross validation techniques. We calculate Net Sentiment Score based on classification results from experiment with best performance. The result shows which company offers better service quality and the best algorithm to classify data.

## III. Method Description

Our study consists of five steps that will be explained below. We use two sentiment analysis tools; R and Rapidminer. We collect data by crawling through Twitter API with R. Then, we use Rapidminer in data pre-processing, classification and classifiers performance evaluation. Fig. 1 describes classification flow in this study.

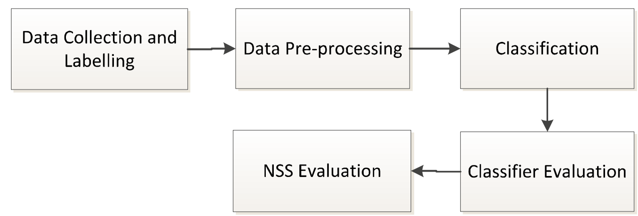


Fig. 1. Classification process

### A. Data Collection and Labelling

We obtained the data using R package twitteR that crawls through Twitter API. Data was gathered every day for a month from February 16, 2016 to March 16, 2016. We presume that customers mention company's account for negative comments while customers express positive comments in non-mention tweets. Data were extracted by keywords "@gojekindonesia", "Gojek", "@GrabID", and "Grab" and we only took tweets in Indonesian. We wanted to see if there was any sentiment difference between mention ("@gojekindonesia" and "@GrabID") and non-mention ("gojek" and "grab") tweets. We obtained 19,918 tweets for "@gojekindonesia", 34,272 tweets for "Gojek", 6,503 tweets for "@GrabID", and 65,712 tweets for "Grab". We only took 2,500 tweets from "Gojek" and "Grab" by random sampling.

Three human annotators, all are information system students, labeled each tweet as positive, negative, neutral, and spam. A simple guideline is given to the annotators to label the tweets. Tweets that contain positive or negative customer's review are labeled as positive or negative, and tweets with questions or general statement are labeled as neutral. If annotators cannot understand the context of a tweet or if tweets are writer by drivers, then tweets are labeled as spam. This manual labelling process also filter tweets with no correlation with this study, especially tweets with "Grab" keywords which is a common word.

After manual labelling, we removed tweets labeled as spam. In the end, we have 9,191 tweets detailed for classification.

We formed two datasets, A and B. Dataset A contains balanced dataset, each sentiment category have same amount of tweets. While dataset B have unbalanced tweets, we use all tweets without balancing the amount of tweets in each category.

### B. Data Pre-Processing

Pre-processing cleans and prepares data for classification [8]. Online texts contain noises and uninformative data that might impact classifiers' performance [8]. Fig. 2 describes pre-processing flow.

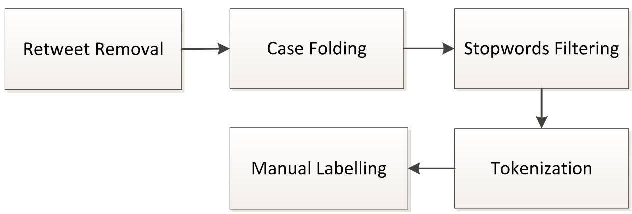


Fig. 2. Pre-processing flow

- Retweet removal
  Same tweets or retweets (RT) are removed. This step deletes tweets that starts with RT.
- Case folding
  All texts are converted into lower case.
- Stopwords filtering
  Stopwords are words that have no particular meaning. Indonesian stopwords, such as "yang", "apa", etc. are removed from texts.
- Tokenization
  Tokenization separates words in a sentence sentence into some parts or tokens. A token contains only one word.
- Manual labelling
  Three human annotators manually labeled each tweet as positive, negative, neutral, or spam. The labelled data will be used to train classifiers.

### C. Classification

We formed classifiers combining different dataset, algorithms, K value in K-fold cross validation, and sampling technique in K-fold cross validation. We classified text with Support Vector Machine, Naïve Bayes, and Decision Tree algorithms. Classification schemes are described in Table I.

TABLE I. CLASSIFICATION SCHEMES

| Case | Dataset | Sampling Method |
|---|---|---|
| I | A | Random sampling |
| II | A | Stratified random sampling |
| III | B | Random sampling |
| IV | B | Stratified random sampling |

Support Vector Machine, Naïve Bayes, and Decision Tree are top machine-learning algorithm in classification [9]. All of them are supervised-learning method that depends on existence of labelled data as training data [10].

Support Vector Machine determines linear line (hyperplane) that separates data into categories by calculating the longest distance between categories of support vectors (data) [10]. Naïve Bayes has class-conditional independence assumption which assumes every known attribute do not affect predicted attribute. Naïve Bayes calculates probability of data belongs to which category using Bayesian theorem. Decision tree classifier forms hierarchical tree from training data to classify testing data.

### D. Classifier Evaluation

In this step, performance of each classifier was tested with the same dataset. Each classifier accuracy, precision, and recall score was measured to select the best classifier to compute NSS. The performance of classifiers is defined by how well the result agrees with human judgement.

### E. NSS Calculation

NSS calculation only relies on labelled or predicted positive and negative tweets. We extracted positive and negative tweets from manually labelled data and classifiers to calculate NSS. We calculate the average of NSS from each company, keyword, and algorithm. NSS has been proven to have the same result with conventional customer satisfaction measurement method. NSS can be any number from -100 and 100, higher score indicates higher customer satisfaction level. We compared NSS score from manually labeled data and classifiers.

## IV. RESULT AND DISCUSSION

### A. Performance Evaluation

There are four experiment variables to be evaluated. They are dataset, classification algorithm, K and sampling technique value in K-fold cross validation.

First, we compared accuracy achieved using dataset A and B. In Table II, we can see that dataset A's highest accuracy score is 61, 11% and dataset B's is 72, 97%. We balanced tweets in each sentiment in dataset A. Balancing the data do not improve the performance of the classifier since the accuracy score is lower than dataset B's. In contrast, experiment with dataset B (unbalanced) has better performance than balanced dataset.

Secondly, the best classification algorithm varies among different classification schemes. Table II shows highest accuracy score of each algorithm. In scheme I and II, Naïve Bayes and Decision Tree have the highest accuracy (61, 11%).

In scheme III and IV, Support Vector Machine and Decision Tree have the same highest accuracy score (72, 97%). We can infer that Naïve Bayes works better in smaller and balance dataset, while Support Vector Machine can perform well in large and unbalanced dataset. Decision Tree can achieve high accuracy in different kind of dataset.

Keep in mind that high accuracy score in some experiments are achieved by grouping tweets into only one sentiment category. Those experiments have low recall and precision score. Therefore, classification results from those experiments are not reliable.

TABLE II. HIGHEST ACCURACY SCORE

| Case | SVM | Naïve Bayes | Decision Tree |
|---|---|---|---|
| **I and II** | 50% | 61,11% | 61,11% |
| **III and IV** | 72,97% | 60,93% | 72,97% |

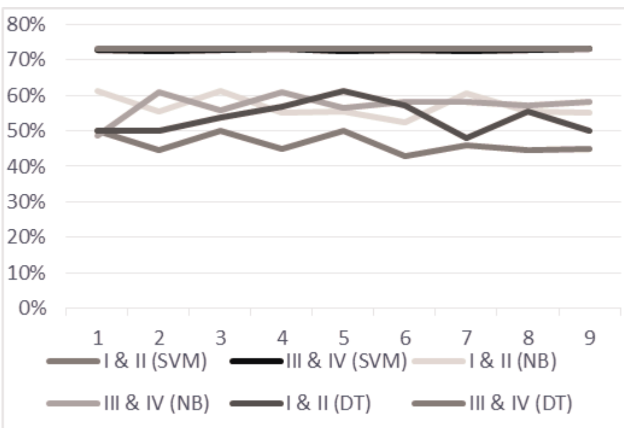


Figure 3 K-values and accuracy in all experiments

Third, we compared K value in K-fold cross validation. K varies between K=2 through K=10. We cannot find any pattern shown in the result. In some cases, highest accuracy scores are achieved in smaller K while in some experiments best K are bigger value. In some experiments, every K achieves the same accuracy score. Figure 3 shows accuracy of various K value.

Last, we used different sampling techniques in K-fold cross validation; random sampling and stratified random sampling. Both techniques gives us same accuracy score. The result either means sampling techniques do not affect performance or Rapidminer does a good job in sampling, giving us same classification result.

### *B. NSS Calculation*

We picked experiments with best performance from every keyword and algorithm to calculate NSS. We compared the score with NSS from manually labeled data. All experiment selected use random sampling in K-fold cross validation since different sampling techniques show same result. NSS formula is:

$$NSS = \frac{Pos.Mentions - Neg.Mentions}{Pos.Mentions + Neg.Mentions} x\ 100 \quad (1)$$

Table III and IV shows the NSS calculation from different classification method.

TABLE III. NSS CALCULATION FROM MANUALLY LABELED DATA

| Company | Keywords | Pos. | Neg. | NSS | Company Avg. |
|---|---|---|---|---|---|
| GO-JEK | "@gojekindonesia" | 251 | 1809 | **-75.63** | **-35.88** |
| | "Gojek" | 121 | 112 | **3.86** | |
| Grab | "@GrabID" | 100 | 487 | **-65.93** | **9.14** |
| | "Grab" | 70 | 6 | **84.21** | |

We can see from Table III that GO-JEK has average NSS -35,88 and Grab has 9,14 based on manually labeled data. From table IV, we can see that GO-JEK has average NSS -15,35 and Grab has 29,33. Based on both classification result, Grab has better NSS than GO-JEK and GO-JEK has both negative score while Grab has both positive score.

Keywords "Gojek" and "Grab" are non-mention tweets. Users do not send the tweets directly to the companies account. Non-mention tweets have higher NSS than mention tweets ("@gojekindonesia" and "@GrabID"). Keywords "Gojek" and "Grab" also has positive score which means decent customer satisfaction.

Mention tweets with "@gojekindonesia" and "@GrabID" keywords both have negative NSS which indicates low customer satisfaction. This is the opposite of non-mention tweets which has positive NSS and good customer satisfaction. We conclude that same company could have different NSS using different keywords.

The gap shows that customers tend to mention the companies' account when complaining or having unpleasant experience, proven by negative NSS on mention keywords. In contrast, customers prefer to express the satisfaction with non-mention tweets to companies account, proven by positive NSS which indicates higher customer satisfaction.

We can see that NSS extracted from manually labeled data is higher than NSS from classifiers. The difference is pretty significant that indicates the classifiers using three different algorithms and classification techniques are not very reliable. The accuracy is far below compared with human judgement.

Keep in mind that some experiments used to calculate NSS do not achieve >70% accuracy which means low or moderate performance. The performance is not very high, also precision and recall score are low. Therefore, the result are not very reliable.

TABLE IV. NSS CALCULATION FROM CLASSIFIERS

| Company | Keywords | Algorithm | Highest Accuracy | Positive Tweets | Negative Tweets | NSS | Average | Company Average |
|---|---|---|---|---|---|---|---|---|
| GO-JEK | "@gojekindonesia" | SVM | 67,68% | 0 | 0 | 0.00 | -7.54 | **-15.35** |
| | | NB | 36,84% | 77 | 122 | -22.61 | | |
| | | DT | 67,68% | 0 | 0 | 0.00 | | |
| | "Gojek" | SVM | 57,44% | 13 | 9 | 18.18 | -23.17 | |
| | | NB | 43,88% | 4 | 61 | -87.69 | | |
| | | DT | 72,97% | 0 | 0 | 0.00 | | |
| Grab | "@GrabID" | SVM | 72,97% | 0 | 0 | 0.00 | -29.23 | **29.33** |
| | | NB | 42% | 4 | 61 | -87.69 | | |
| | | DT | 72,97% | 0 | 0 | 0.00 | | |
| | "Grab" | SVM | 71,28% | 61 | 0 | 100.00 | 87.89 | |
| | | NB | 61,25% | 5 | 1 | 66.67 | | |
| | | DT | 64,29% | 66 | 1 | 97.01 | | |

## V. CONCLUSION AND FUTURE WORK

Based on classification and NSS calculation results, Grab has higher NSS than GO-JEK meaning Grab's customer satisfaction level is higher than GO-JEK's. Support Vector Machine and Decision Tree achieves highest accuracy score.

Classifiers achieve better performance in larger and unbalanced dataset. Different algorithms works best in different kind of dataset, Naïve Bayes works well in small dataset while Support Vector Machine performs best in large dataset. Decision Tree performs well in both dataset. High accuracy score is not followed by high precision and recall score, since in some experiments classifiers tend to group data only into one sentiment. We cannot see any pattern in K value in K-fold cross validation. Different sampling techniques in K-fold cross validation do not affect classifier performance.

For future research, we suggest to use more specific keywords to collect data. For example, use product name instead of company name, GO-RIDE and GO-CAR for GO-JEK and GrabBike and GrabCar for Grab. We also advice to use features in pre-processing and classifying data to form better classifier, thus get better classification result.

## REFERENCES

[1] S. Neti, "Social Media and Its Role in Marketing," *International Journal of Enterprise Computing and Business System,* vol. 1, no. 2, pp. 1-15, 2011.

[2] N. A. Vidya, M. I. Fanany and I. Budi, "Twitter Sentiment to Analyze Net Brand Reputation," *Procedia Computer Science 72,* pp. 519-526, 2015.

[3] L. Zhang, K. Huac, H. Wang, G. Qiane and L. Zhang, "Sentiment Analysis on Reviews of Mobile Users," *Proceedings of The 11th International Conference on Mobile Systems and Pervasive Computing,* pp. 458-465, 2014.

[4] Y. Zhang and P. Desouza, "Enhance The Power of Sentiment Analysis," *International Journal of Computer, Electrical, Automation, Control, and Information Engineering,* pp. 421-426, 2014.

[5] M. Stelzner, "2015 Social Media Marketing Industry Report," May 2015. [Online]. Available: http://www.socialmediaexaminer.com.

[6] W. Chamlertwat, P. Bhattarakosol and T. Rungkasiri, "Discovering Customer Insight from Twitter via Sentiment Analysis," *Journal of Universal Computer Science,* pp. 973-992, 2012.

[7] E. Bricker, "Can Social Media Measure Customer Satisfaction?," 2011. [Online]. Available: http://www.netbase.com/.

[8] X. L. Y. S. Emma Hadi, "The Role of Text Pre-processing in Sentiment Analysis," *Procedia Computer Science,* vol. 17, pp. 26-32, 2013.

[9] X. Wu et al., "Top 10 Algorithms in Data Mining", *Knowledge and Information Systems,* vol. 14, pp. 1-37, 2007.

[10] W. Medhat, A. Hassan, H. Korashy, "Sentiment Analysis Algorithms and Applications : A Survey", *Ain Shams Engineering Journal,* pp,1093-1113

Topic
Machine Learning and Computer Vision