Sentiment Analysis on Twitter Posts: An analysis of Positive or Negative Opinion on GoJek

Ike Pertiwi Windasari
Dept. of Computer System
Diponegoro University
Semarang, Indonesia

Fajar Nurul Uzzi
Dept. of Computer System
Diponegoro University
Semarang, Indonesia

Kodrat Iman Satoto
Dept. of Computer System
Diponegoro University
Semarang, Indonesia

Abstract— **Online transportation, such as GoJek, is preferred by many users especially in areas where public transport is difficult to access or when there is a traffic jam. Twitter is a popular social networking site in Indonesia that can generate informations from users' tweets. In this study, we proposed a system that detect public sentiments based on Twitter post about online transportation services especially GoJek. The system will collect tweets, analyze the tweets sentiments using SVM, and group them into positive and negative sentiment.**

Keywords— Support Vector Machine (SVM), GoJek, Tweets

I. INTRODUCTION

The development of information and communication technology makes it easier for people to connect with each other and share information, one of them is through social networking. Social networking media is an online medium where one can propose an idea and thoughts to a particular problem or share things that are considered interesting both through digital and multimedia text. One of the most used social networking in Indonesia is Twitter.

Twitter allows its users to write various topics and discuss issues that occur. Unlike some other social media that require the consent of both parties to connect to each other, Twitter allows users to keep track of submissions (called tweets) from other users without getting approval. This is one of the reasons why Twitter is the place for the flow of information.

Millions of posts from twitter can be used to extract useful information in various fields. One of the new information that can be obtained by analyzing Twitter data is people's sentiment towards a topic. By knowing the public sentiments, problem can be evaluated for improvement.

Online transport services is one of the topics that is discussed a lot on Twitter. Many people rely on online transportation, both the drivers and customers. Online transportation has become a necessity especially as public transportation in some areas are difficult to reach. In addition, online transportation is also popular in areas hit by traffic jams. Public response to the services provided by online transportation service providers are varied, there is a positive and negative respond. By knowing the sentiments, users can find out the popular and well accepted online transport providers. The service provider can also use the information to analyze public responses to improve the quality of its services. GoJek is a local online transportation provider that is very popular in Indonesia. The services provided by GoJek are varied, not only transportation services but also food purchase and delivery services.

Public sentiment information can be obtained by observing the tweets. But to collect and observe millions of tweets will require a lot of time and effort. Therefore, the existence of an application that automatically crawl and analyze the twitter sentiments will be very useful in information extraction.

In this paper, Twitter dataset is used and analyzed using unigram and TF-IDF feature extraction technique. Further, The dataset is classified using Support Vector Machine (SVM) and generate a positive and negative detection on Twitter post. Thus, this study intend to create system that can automatically detect public sentiments based on Twitter post about online transportation services especially GoJek. This system are expected to collect tweets, analyze the tweets sentiments and group them into positive and negative.

II. RELATED WORK

Existing researches in the opinion mining and sentiment analysis are mentioned below. Ariwibowo [1] in his research made the design of Twitter opinion mining architecture using Case Based Reasoning method where it was concluded that opinion mining requires preprocessing stage. The result of the research is the drawing of the architectural design application of the sentiment analysis toward the product brand. Hidayatullah [2] compared Naïve Bayes and Support Vector Machine classification method with TF-IDF feature extraction methods. SVM method produces better performance accuracy compared to Naïve Bayes, but both methods already have good results for the tweets classification.

Other research in sentiment analysis by Bhadane [3] focuses on the various methods used for classifying a given piece of natural language text according to the opinions expressed in it i.e. whether the general attitude is negative or positive. They implemented a set of techniques for aspect classification and polarity identification of product review using machine learning (SVM) combined with domain specific lexicons. Their experimental results indicate that the proposed techniques have achieved about 78% accuracy and are very promising in performing the tasks.

Balahur [4] in her research presented a method to classify the tweets sentiment by taking the peculiarities and adapting the features employed to their structure and content. She employed a pre-processing stage to normalize the language and generalize the vocabulary employed to express sentiment. With that approach the system can obtain good result even tough using minimal linguistic processing. The use of such generalized features significantly improves the results of the sentiment classification, when compared to the best performing approaches that do not use affect dictionaries. Han [5] in his research combine SVM classifier and character N-gram language models for sentiment analysis on Twitter text. Features derived from character n-gram language models do not perform very well, but they may benefit from a larger training data set.

III. Working Procedures

In this paper, the sentiment analysis of Twitter post with GoJek related keyword is obtained through several processes that is tweet crawling, manual labeling for training dataset, text preprocessing, feature extraction, and SVM classification that is shown in Fig 1. The SVM classification in this system is divided into data training and data prediction.

A. *Tweet Crawling*

Dataset (tweet) was collected from Twitter using Twitter API with keywords related to GoJek, such as GoJek, GoFood,and GoCar.

B. *Manual Labeling*

Manual labeling is given for each tweet data that has been crawled. Manual labeling is used for training data on SVM classification.

C. *Text Preprocessing*

Text preprocessing is the process of preparing the initial text data into easier data to be further processed into basic text form and eliminating the noise. This stage aims to obtain more optimal calculation results. The steps taken in preprocessing text in this research are the convert emoticon, cleansing, case folding, stemming, convert negation, stopword removal, and tokenizing.

D. *Feature Extraction*

The n-gram form for the term used in this paper is word-based unigram. The word-based unigram form means the system will process the weighting of n-grams on every single word that makes up the tweets. The tweet which contains more rare words that have a higher weight than which contain common words and it has made a greater effect on the classification task.

Term weighting is assigning value to each term derived from the training tweets. The scoring aims to find out how important a term in representing a sentence. The term weighting method used in this study is Term Frequency - Inverse Document Frequency (TF-IDF) method. The determination of term score in the TF-IDF method is based on the frequency of occurrence of terms in a document.

The TF-IDF value of the term can be found using the following equation:

$$Wi = TF(\omega i, d) \times IDF(\omega i) \quad (1)$$

where *Wi* is the weight of word *ωi* in document *d∈D, TF(ωi, d)* is the term frequency, or the number of *ωi* in *d*, and

$$IDF(\omega i) = \log(|D|/DF(\omega i)) \quad (2)$$

where *IDF* is the inverse document frequency and *DF*(*ωi*) represents the appearance of ωi in D. The largest value of *IDF*(*ωi*) occurs when *ωi* appears only in one document and its effect is particularly substantial.

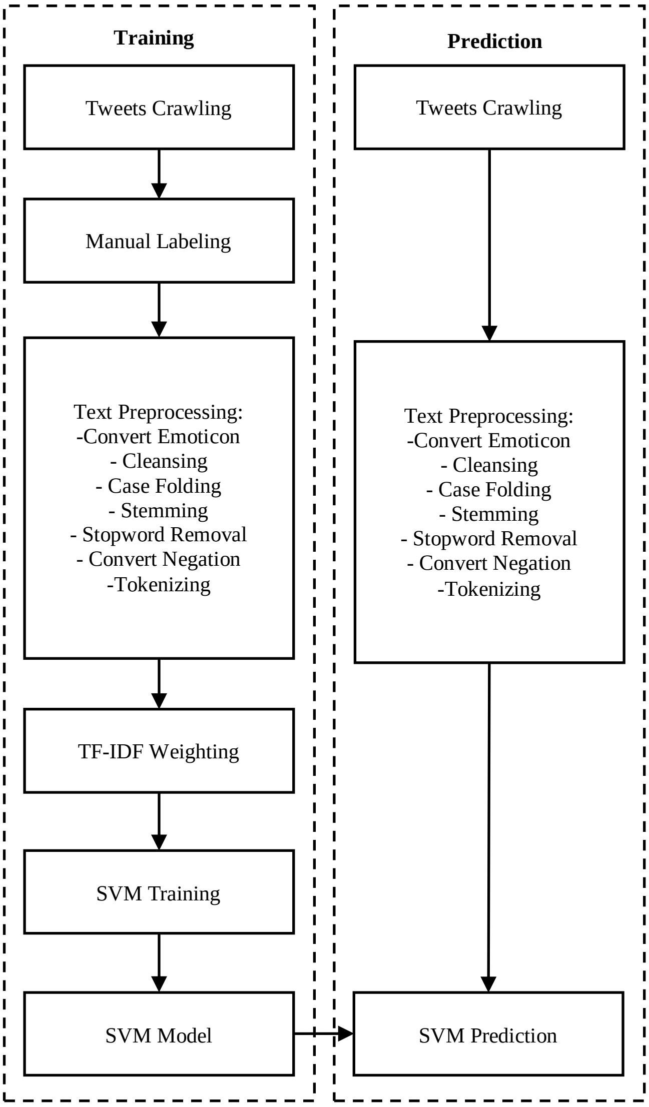

Fig. 1. Sentiment Analysis

E. *SVM Classification*

Support Vector Machine (SVM) in this research is used in both training and stage. The use of SVM in specific training for the model to be used as a reference in the access stage, while the use of SVM in this stage is already known class. In this research, there are two classes that are positive class and negative class. Every tweets included in the positive class have the label "1" while the negative class has the label "-1". This study uses LibSVM for PHP by Ian Barber. LibSVM is a library that supports SVM functions such as training and classification. The input data format for SVM processing is an array.

IV. Result and Discussion

The system has gathered tweets in Indonesian language as a training data, which consisting of 1000 positive tweets and 1000 negative tweets. Text preprocessing in the system is done with order as shown in Fig.1.

We convert the emoticons to prevent it from being deleted in the cleansing process, cleanse the text from punctuation, case folding to change all case to lower case, stemming text to change it into basic form, stopword removal, change negation, and tokenizing sentence into terms.

We put the stemming process before stopword removal with the aim of making the word "adakah" to be the word "ada" so that the word will still be removed from the tweet text.

The process of convert negation is done after stemming and stopword removal with the purpose of making the sentence "tidak ada yang mengantar" to "xantar". Tokenizing breaks the sentence into a word where the word is used as a term. The text preprocessing is shown in Table 1.

TABLE I. Text Preprocessing

No	Teks awal	*Text preprocessing*	*Tokenizing*
1.	Makasih gojek di saat saya lapar seperti ini kamu selalu menolongku.	makasih gojek lapar tolong	*Term* 1 : makasih *Term* 2 : gojek *Term* 3 : lapar *Term* 4 : tolong
2.	Ini udah ga ada abang gojek yah di Bandung? Susah dapatnya.	udah xada abang gojek bandung susah dapat	*Term* 1 : udah *Term* 2 : xada *Term* 3 : abang *Term* 4 : gojek *Term* 5 : bandung *Term* 6 : susah *Term* 7 : dapat

Terms weighting using TF-IDF method is shown in Table 2. Below is two example of tweets that will put into text preprocessing and TF-IDF weighting.

1. Enaknya kalau dapet driver gojek yang enak diajak ngobrol.
2. Sedihnya pesen gojek gak nemu-nemu drivernya.

After going through the text preprocessing stage, tweets sentences turned into text as follows:

1. Enak dapet driver gojek enak ajak ngobrol
2. Sedih pesen gojek xnemu nemu driver.

TABLE II. TF-IDF Weighting

No	*Term*	*Tf*		*Df*	*Idf*	*Tf-idf*	
		T1	*T2*			*T1*	*T2*
1	Enak	2	0	1	0,30102	0,60204	0
2	Dapet	1	1	2	0	0	0
3	Driver	1	1	1	0,30102	0,30102	0,30102
4	Gojek	1	1	2	0	0	0
5	Asyik	1	0	1	0,30102	0,30102	0
6	Ajak	1	0	1	0,30102	0,30102	0
7	Ngobrol	1	0	1	0,30102	0,30102	0
8	Sedih	0	1	1	0,30102	0	0,30102
9	Pesen	0	1	1	0,30102	0	0,30102
10	Xnemu	0	1	1	0,30102	0	0,30102
11	Nemu	0	1	1	0,30102	0	0,30102

V. Conclusion

Millions of Twitter users post their opinion on their tweets. Business can use this information to their advantage, but it takes a lot of time. Therefore, there is a need of sentiment analysis that predicted tweet sentiment towards a topic. In this research, we limit our keyword related to online transportation especially GoJek.

We preprocess the text obtained from Tweets to eliminate the noise, then we use n-gram unigram and TF-IDF as feature extraction methods and use SVM algorithm for classification methods. The result is a prediction of Tweets data that is considered to be a positive or negative sentiment towards GoJek service. From the test results obtained an accuracy of 86%, prediction error rate of 14%, the correct prediction rate for 100% positive sentiment, and the correct prediction rate for negative sentiment 67.44%.

In the future, we would study by performing larger scale experiments by having larger data set. We would also want to compare the result using other classification method and feature to improve the accuracy.

References

[1] G. Eason, B. Noble, and I.N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (*references*)

[2] Aribowo, A. S., “Arsitektur Aplikasi Twitter Opinion Mining Untuk Mengetahui Sentimen Publik Terhadap Merek,” pp. 14–20, November 2015.

[3] Balahur, A., “Sentiment Analysis in Social Media Texts,” pp. 120–128, June 2013.

[4] Bhadane, C., Dalal, H., & Doshi, H., “Sentiment analysis: Measuring opinions,” Procedia - Procedia Computer Science, *45*, pp. 808–814. 2015.

[5] Han, Q., & Guo, J. (n.d.), “CodeX : Combining an SVM Classifier and Character N-gram Language Models for Sentiment Analysis on Twitter Text.

[6] Hidayatullah, A. F., & SN, A. Analisis sentimen dan klasifikasi kategori terhadap tokoh publik pada twitter,” pp. 1–8. 2014.