OPEN ACCESS
IJOICT
ISSN 2356-5462
socj.telkomuniversity.ac.id/ijoict

Intl. Journal on ICT
Vol. 2, Issue. 1, June 2016. pp. 57-66
doi:10.21108/ijoict.2016.2.1.85

Sentiment Analysis on Twitter about the Use of City Public Transportation Using Support Vector Machine Method

Veronikha Effendy [#1], Anita Novantirani [*2], Mira Kania Sabariah [#3]

 # Telkom Schools of Computing, Telkom University
Telekomunikasi Terusan Buah Batu Street, Bandung

[1] veffendy@telkomuniversity.ac.id
[2] anitnitnit@gmail.com
[3] mirakania @telkomuniversity.ac.id

Abstract

Traffic jams that occur in big cities in Indonesia due to the increased use of private vehicles. One solution to overcome this problem is to increase the use of public transport. But, the existing public transport is still not much in demand by the community. Some people express their opinions regarding the use of city public transportation via Twitter. The opinions can be processed as a sentiment analysis to determine the positive opinions and negative opinions. The opinion will then be analyzed to determine factors that are the main cause of the ineligibility use of public transport as well as the factors that make the public choose to use this type of transport. By upgrading of facilities and services based on the results of sentiment analysis, it is expected that people will switch to use city public transportation, which would reduce the traffic jam. This research used SVM method to process sentiment analysis. The result has shown SVM accuracy reaches 78.12%, which indicates that the results of this reserach deserve to be considered.

Keywords: traffic jam, sentiment analysis, SVM

I. Introduction

Traffic jam often occurs in various region, particularly in big city area in Indonesia. Head of the Jakarta City Transportation Council, Azas Tigor Nainggolan, assumed that the congestion is caused by increase the use of private vehicles (Irmansyah, 2013). Increased use of public transport can be a key solution for congestion problems (Furqan & Sjafruddin, 2013), which if successful would make private vehicle users will switch to public transport and will reduce the amount of volume of vehicles on the road which would reduce congestion. However, most people are still reluctant to use public transport and prefer to use private vehicles. The reluctance is based on several factors, among others, the travel time, cost, safety and security, as well as the pleasure and convenience of the users of public transport itself (Manheim, 1979). This needs to be traced further to determine public opinion about what is felt when using city public transportation in Indonesia.

Opinion regarding the use of public transport can be obtained using one of social media, namely Twitter. Twitter is a familiar medium used by the Indonesian people to express their impression and feelings about the use of public transport. Ease of access Twitter in the delivery of opinion could be an opportunity to be used as assessment and evaluation of city public transport services in Indonesia. To generate information from the existing opinion data, the data is processed with sentiment analysis that will separate opinion in a positive or negative sentiment class, and infer what factors are often discussed in those opinions. Previous research has been done on the Twitter sentiment analysis with various methods such as Naive Bayes Classification, Maximum Entropy, or Support Vector Machine (Nur & Santika, 2011) (Go, Bhayani, & Huang, 2009). It was concluded that the use of Support Vector Machine (SVM) provides superior results compared to other

Received on Jan 2016. Accepted on June 2016

methods that the level of accuracy is up to 82.2% (Go, Bhayani, & Huang, 2009). Moreover, SVM method itself can be used to classify the data based on attribute valuation opinion held to be separated opinions belong to a class of positive or negative. In another study also stated the use of proven SVM can classify sentiment data in Indonesian language with the highest accuracy, reaching 92.5% compared with the use Naive Bayes with 90% accuracy and C4.5 with an accuracy of 77.5% (Wulandini & Nugroho, 2009). For these reasons, this research used the SVM method as a method of classification.

II. LITERATURE REVIEW

A. Twitter

Twitter is a website owned and operated by Twitter Inc., which offers a microblog social networking (Zhang, Ghosh, Dekhil, Hsu, & Liu, 2011). Called microblog, because the site allows its users to send and read messages such blogs in general but limited to a number of 140 characters displayed on the user profile page. Twitter has a characteristic and unique writing format with symbols or special rules. Messages in Twitter known as tweets.

B. City Public Transportation

City public transport, also called public transportation, is transport from one place to another within the city by bus or public passenger cars involved in fix and regularly route (Direktur Jenderal Perhubungan Darat, 2002).

Availability of public transport will certainly provide benefits to society in general. Society will be helped if you want to travel to a destination. If public transport is maintained and cared for, it will give good impact, which is the public would prefer to use public transport than private vehicles.

C. Support Vector Machine

Support Vector Machine (SVM) is one of classification method using machine learning (supervised learning) that predict class based on the model or pattern of the results of training process. Classification is done by finding the hyperplane or boundary line (decision boundary) that separates between classes. SVM search hyperplane value by using support vector and the value of margin (Han & Kamber, 2006). Separator function or hyperplane is a linear function in equation (1).

$$w.x+b=0 \quad (1)$$

In equation (1), w is the weight that present the position of hyperplane in normal field, x is the vector of input data, and b is the bias that represents the position of the field relative to the origin.

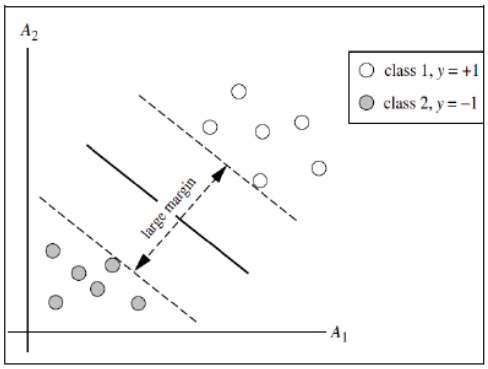

Figure 1 SVM with the best hyperplane (Han & Kamber, 2006)

The best hyperplane function is one that has the largest margin value, where the margin is the distance between the hyperplane to the nearest point of each class. The closest point is called a support vector. Figure 1 shows SVM with the best hyperplane (Han & Kamber, 2006).

III. RESEARCH METHOD

A. Data

The dataset used in this research is an Indonesian-language opinion about city public transportation from Twitter. Dataset form twitter opinion which is basically a public expression of a case through social media. In social media Twitter, users tend to use non-formal language, without using proper grammar and appeared many slang words.

Dataset focus on Indonesian opinion that discusses the city public transport by selecting the four sample vehicle with details number of dataset shown in table 1.

TABLE I
DETAIL NUMBER OF DATASET

Dataset	Details of the Data
Angkot	272 opinions
Kopaja	184 opinions
Metro Mini	264 opinions
Transjakarta	481 opinions

B. Preprocessing

Stages in the process of preprocessing in this research are as follows:

1) Data Cleansing: The process consists of case folding and removes noise. Noise in this case is a character other than letters (numbers, symbols, and punctuation).
2) Tokenization: The process of cutting a row of words in the document into a single word pieces.

3) Word Normalizes: The process of converting not standard word to a standard word conducted with reference to the standardization word dictionary derived from previous research (Wicaksono & Purwarianti, 2010). In this research, the dictionary was developed by modifying and adding handling adjusted to the dataset situation.
4) Part of Speech (POS) Tagging: The process of tagging the word. In this research, the process of tagging using HMM word and Rule Based POS Tagging.
5) Stop word Removal: Stop word removal process, that is words that often appear but do not have a specific meaning and is not considered important in the opinion classification.

C. Weighting

Weighting process is performed based on the number of occurrences of words in a document, so that the document can be represented in a vector. Feature weighting used is unigram, and Term Frequency-Inverse Document Frequency (TF-IDF) weighting method.

D. Support Vector Machine (SVM) Classification Method

The process of classification in this research use Support Vector Machine (SVM) with linear kernel. SVM is a classification method that predicts the class based on model or pattern of the results of the training process. Classification is done by finding the hyperplane using the formula in equation (1) or a boundary line (decision boundary) that separates between a class with another class, which in this case is the line separate positive tweets (labeled +1) with negative tweets (labeled -1). SVM search hyperplane value by using support vector and value of the margin (Han & Kamber, 2006). In this research, the data input which has a vector representation is obtained from the weighting process. By doing training process in SVM classification, it will generate a value or pattern to be used in the testing process for testing the SVM, which aims to label sentiment class in tweets (Pang, Lee, & Vaithyanathan, 2002).

E. Validation and Evaluation

Validation is the act of proving that a process or method gives consistent results and achieve the results according to specific and well-documented. The validation process in this research using the k-fold cross validation with k=2 up to k=10.

The performance evaluation was conducted to test the results of classification by measuring the value of the performance that performed by the system. Testing parameters that is used to evaluate is accuracy or truth level of classification process which calculation obtained from the coincidence matrix table.

IV. RESULTS AND DISCUSSION

System testing performed on each dataset separated by types of public transport.

A. Testing Scenario

Testing in this research is performed on dataset using the parameter of different composition of training and testing data, and the parameters of the different composition of positive and negative data.

1) Testing Dataset Scenario with Parameter of Different Composition of Training and Testing Data

Testing is done by dividing the training data and testing data with a composition based on the partition of data on k-fold cross validation with k=2 up to k=10. This test aims to determine the most optimal k-fold value in the opinion classification process of this case research.

2) Testing Dataset Scenario with Parameter of Different Composition of Positive and Negative Data

Testing is done by using a number of documents that are not balanced between positive data and negative data, ie the number of documents more positive than negative document, and vice versa. Tests carried out several times by randomize the dataset with the intention to recognizing fairness level of the data test result.

In research (Zhang, Weishi, Ding, Chen, & Li, 2010), it is said that the composition of the positive and negative of data need to be balanced to avoid any bias or weight deflection with a high number of occurrences in negative data only or positive data only. Therefore this test done to prove whether the situation also applies to the data of this research. This test will be performed on the data that have the best accuracy from the test results of scenario 1.

3) Opinion Factor Testing

This test is done to determine the most dominating opinion of each type of public transport dataset. Testing is done by using the word class labels from POS tagging results, then performed opinion extraction by selecting words which have word class adjective, noun, adverb, and verb. From many factors of the opinions, taken ten most.

B. Analysis of Test Results

This section is shown the results of the testing and analysis of the results that have been obtained from the testing process.

1) Accuracy Analysis with Different Composition Training and Testing Data

After testing by using the composition of training data and test data differently based on the partition of data on k-fold cross validation, the resulting accuracy is shown in table II. From these test it can be seen the composition of training data and test data provide the most optimal accuracy results. On Angkot dataset, the optimal composition data i.e. training data :testing data is 3:1 (4-fold), while for the dataset Kopaja, Metro Mini, and Transjakarta found that the optimal composition is 1:1 (2-fold).

TABLE II
ACCURACY VALUE BASED ON THE K-FOLD

Fold	**Type of Public Transportation**			
	Angkot	***Kopaja***	***Metro Mini***	***Transjakarta***
2	70.9559	**65.7609**	**71.9697**	**78.1250**
3	69.5249	63.0466	67.8030	73.2677
4	**72.0588**	64.6739	67.0454	76.6827
5	66.9709	58.2748	67.7778	72.1429
6	66.8808	59.7569	69.6969	73.5154
7	67.6128	61.4228	69.4235	74.9746
8	69.4853	63.5417	63.9476	75.7211
9	63.7037	58.1818	65.1323	74.2854
10	64.0934	59.7778	61.0714	73.6071

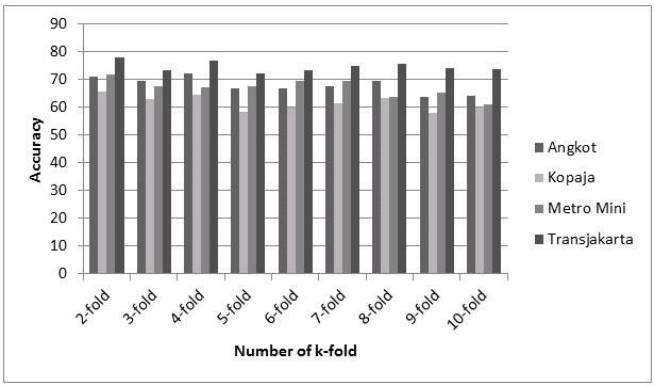

Figure 2 Graphics of Accuracy Value based on K-fold

The resulting accuracy of the classification model is influenced by the proportion of training data and test data. As seen in table II and figure 2, greater the fold value or more the amount of training data, the accuracy value result tend to decrease.

Based on test results, Trans Jakarta dataset produces better accuracy than other dataset, with the accuracy reach 78.125%. This is because the dataset Trans Jakarta have higher numbers than other datasets so that the higher the diversity of data that affect the models generated from the training process. Instead, Kopaja dataset that has the least amount of data has the smallest accuracy compared to three other datasets.

The accuracy results in this research are also influenced by the characteristics of the dataset, the dataset derived from Twitter and using Indonesian language. Indonesian language has a grammar that is more complex than English grammar which has special rules for writing a particular sentence patterns. In addition, the data tend to use grammar Twitter irregular and there are many non-standard words. Therefore the use of the word in the dictionary for the normalization of this research is also an important influence on the accuracy results.

2) Accuracy Analysis with Different Composition Positive and Negative Data

The data will be used in this testing scenario is data that has the best accuracy results on testing scenario 1, i.e. the Trans Jakarta dataset. In this test, the dataset is divided into three types with details as in table III.

TABLE III
THE DIVISION OF DATASET

No.	Dataset Name	Number of Tweet	Positive Tweet	Negative Tweet
1	Dataset A	350	200	150
2	Dataset B	350	150	200
3	Dataset C	350	175	175

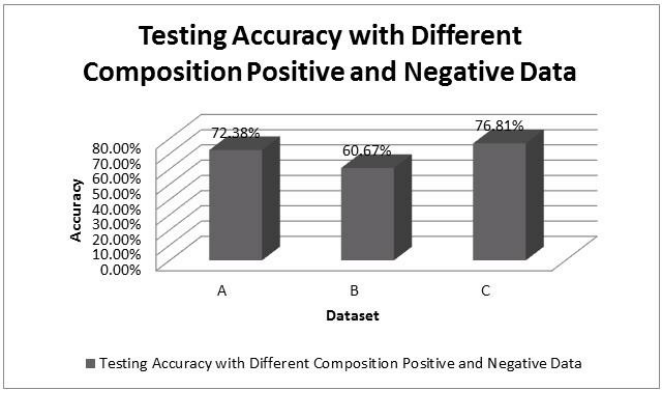

Figure 3 Accuracy value based on the amount of positive and negative data

Figure 3 shows that the accuracy of the dataset C has the best accuracy values among the three existing datasets, amounting to 76.81%. In accordance with what is disclosed in the previous research (Zhang, Weishi, Ding, Chen, & Li, 2010) that the dataset C which has a balance composition of positive and negative data will result in a better classification since avoiding bias or deflection of weight word with a high number of occurrences in the negative data only or positive data only.

3) Opinion Factor Analysis

After testing for opinion factor using opinion extraction, the obtained results of dominant factors based on each type of public transport which are shown in table IV.

TABLE IV
THE DIVISION OF DATASET

No.	Type of Public Transportation	Word	Frequency	Type of Sentiment
1	Angkot	Mengetem	32	Negative
2		Keren	25	Positive
3		Tidak tersedia	22	Negative
4		Sialan	19	Negative
5		Menunggu	18	Negative
6		Lama	17	Negative
7		Macet	17	Negative
8		Hemat	16	Positive
9		Enak	15	Positive
10		Rokok	15	Negative
11	Kopaja	Hemat	12	Positive
12		Macet	12	Negative
13		Menunggu	11	Negative
14		Keren	10	Positive
15		Duduk	9	Positive
16		Copet	8	Negative
17		Mogok	6	Negative
18		Penuh	6	Negative
19		Rokok	6	Negative
20		Tidak tersedia	6	Negative

21	Metro Mini	Sembrono	30	Negative
22		Baru	28	Positive
23		Macet	21	Negative
24		Baik	16	Positive
25		Bagus	14	Positive
26		Keren	13	Positive
27		Nyaman	13	Positive
28		Senang	11	Positive
29		Copet	10	Negative
30		Tabrak	10	Negative
31	Transjakarta	Baru	43	Positive
32		Lama (kedatangan)	43	Negative
33		Dingin	41	Positive
34		Menunggu	41	Negative
35		Cepat	38	Positive
36		Mogok	37	Negative
37		Nyaman	23	Positive
38		Sepi	23	Positive
39		Duduk	20	Positive
40		Kosong	20	Positive

Factors underlying the reluctance of people to switch to public transport including travel time, cost, safety and security, and the pleasure and convenience of public transport itself (Manheim, 1979). Based on the test results as shown in table 4, obtained opinions factor according to the statement, which is about the travel time tend to be long and jammed a lot of complaints on Angkot, Kopaja, and Metro Mini, but some are of the opinion faster travel time, especially on Transjakarta. It is also related to the time that is the accuracy of the arrival time of transport, is still a problem because there are many people who complain about the waiting time and „mengetem" (stop long at any place). As for the cost factor is not so dominant appear in the results, it shows that people can receive a range of costs that exist today because it is not a lot of complaints. Safety and security factors are opinions of levity or reckless, especially in transport Metro Mini. In addition, the safety factor is also of concern is the pickpockets or theft, it is complained especially in the transport Metro Mini and Kopaja. Pleasure and convenience factors of the users of public transport is also be a concern. In Transjakarta seen that many users are happy because some things related to pleasure and comfort, which are opinion factors of new vehicle, comfortable, cool, quiet, and easy to get a seat. For Metro Mini opinion factors included good, nice, cool, comfortable. Opinion factor about comfort received for Kopaja is relatively balanced between getting a seat and a full vehicle.

Based on the analysis described above, it can be concluded that the city public transportation have a variety of assessment in the people"s view. There are positive opinions, and there are also negative opinions for each type of transport researched. The positives opinions can be utilized to be able to improved or can be applied to other types of transportation, as well as for negative opinions could be used as advice for stakeholders to be able to improve public transport services, especially city public transport. Improvement of public transport service is expected to make the public want to switch from using private vehicles to be using public transport, which if can be achieved also expected to help reduce congestion

V. CONCLUSION

Based on the results of this research, it can be concluded that sentiment analysis on Twitter data about the use of city public transport can be done using Support Vector Machine, with accuracy reaches 78.12% on

Transjakarta dataset. The accuracy of the results on the use of Support Vector Machine method is influenced by four factors: (1) the composition of the training data and testing, (2) the amount of the dataset used, (3) the composition of the positive data and negative, and (4) use of dictionary word normalization.

This research can be used as a recommendation to increase performance of city public transport. For advice in the future, this research can be developed so that the accuracy results can be improved and can enrich recommendations with the use of much more dataset and use a dictionary that is always updated in accordance with the growth of language.

References

Direktur Jenderal Perhubungan Darat. (2002). Pedoman Teknis Penyelenggaraan Angkutan Penumpang Umum di Wilayah Perkotaan Dalam Trayek Tetap dan Teratur.

Furqan, & Sjafruddin, A. (2013, April 29). Retrieved November 1, 2013, from Prof. Ade Sjafruddin: Angkutan Umum, Solusi Kunci Kemacetan Jakarta: http://www.itb.ac.id/news/3899.xhtml

Go, A., Bhayani, R., & Huang, L. (2009). *Twitter Sentiment Classification using Distant Supervision.* Project Report, Stanford.

Han, J., & Kamber, M. (2006). *Data Mining Concept and Technique.* San Francisco: Elsevier.

Irmansyah, A. (2013, Augustus 20). *PortalKBR.com*. Retrieved November 1, 2013, from Benahi Angkutan Umum Segera agar Tahun Depan Jakarta Tak Macet Total: http://www.portalkbr.com/berita/perbincangan/2897988_4215.html

Manheim, L. (1979). *Fundamental Transportation Systems Analysis* (Vols. I, Basic Concept). The MIT Press.

Nur, M. Y., & Santika, D. D. (2011). Analisis Sentimen pada Dokumen Berbahasa Indonesia dengan Pendekatan Support Vector Machine. *Konferensi Nasional Sistem dan Informatika.* Bali.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing 10*, (pp. 79-86).

Wicaksono, A. F., & Purwarianti, A. (2010, Agustus). HMM Based Part-of-Speech Tagger for Bahasa Indonesia. *Proceedings of 4th International MALINDO (Malay – Indonesian Language)*.

Wulandini, F., & Nugroho, A. S. (2009). Text Classification Using Support Vector Machine for Webmining Based Spation Temporal Analysis of the Spread of Tropical Diseases. *International Conference on Rural Information and Communication Technology*, (pp. 189-192).

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2011, Juni). Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis. *HPL-2011-89, 89*.

Zhang, Weishi, Ding, G., Chen, L., & Li, C. (2010). Chinese Online Video Recomendation by Using Virtual Rating Predicted by Review Sentiment Classification. *IEEE International Conference on Data Mining Workshop*.