Sentiment Analysis on Twitter Using the Combination of Lexicon-Based and Support Vector Machine for Assessing the Performance of a Television Program

Tiara[1], Mira Kania Sabariah[2], Veronikha Effendy[3]
[1, 2, 3] School of Computing, Telkom University
Bandung, Indonesia
[1]araait@students.telkomuniversity.ac.id, [2]mirakania@telkomuniversity.ac.id, [3]veffendy@telkomuniversity.ac.id

Abstract— **The development of social media, especially twitter is growing rapidly. Twitter is usually used to comment on a product, a person or even a television program. The written comments by twitter users can reach hundred thousand or even millions every day. By using the comments obtained from twitter, it can complement a television program assessment that usually done by using rating, which only represented in terms of quantity. Therefore, by analyzing the comments on twitter hopefully it would be able to complement the assessment in terms of quality. The comments from twitter can be analyzed by performing a sentiment analysis process. The methods used in this study are the combination of lexicon-based method and Support Vector Machine. The results show that these combination methods can be implemented in analyzing sentiment on the television program with the accuracy rate that reaches 80%. This value is not influenced by the ratio of training data and test data that are used. However, in this study the data tweets that are dominant by positive sentiment tends to have a higher accuracy rate than the data tweet that consist a balanced amount of sentiment or the one that are dominant by the negative sentiment**.

Keywords— twitter, television program, sentiment analysis, lexicon-based, support vector machine

I. INTRODUCTION

Television is a mass media that most familiar and influential in the society. During this time, the decision whether a television program will be extended or not depends on the quantity, which is the value of rating and share. But the problem is a television program that gets a high rating does not always have a high quality for their show according to the audience perspective. Whereas, not all qualified impressions have a high rating [1].

In recent years, we have witnessed that opinionated postings in social media (e.g., reviews, forum discussions, blogs, micro-blogs, Twitter, comments, and postings in social network sites) have helped reshape businesses, and sway public sentiments and emotions, which have profoundly impacted on our social and political systems [2]. There is a field of study that analyzes this kind of opinionated postings which usually called as sentiment analysis or opinion mining [2]. So by using sentiment analysis could help to complement the assessment for television program in terms of quality by using the media social, Twitter.

The dataset used in this research is a collection of tweets. Those tweets will be extracted and processed so it can produce information such as sentiment that consists in tweets. Sentiment analysis on tweets is used to find out whether a tweet consists of positive or negative sentiment. There are two kinds of learning that usually used in the process of sentiment analysis, which is supervised learning and unsupervised learning [3]. The machine learning method is belong into supervised learning, this method usually need a lot if training data that have been labelled manually. Without labelled training data, supervised learning will not able to be processed [3][4]. Whereas, the lexicon-based method is belong to unsupervised learning, which does not need training data and only depend on the dictionary that is used [3]. Both methods have different characteristics, but it can complementary if both methods are combined. The combination of both methods can be done by using lexicon-based method to create labelled tweets which can be used as training data in Support Vector Machine method so there will be no manual labelling process in this combination methods[5][3].

This combination method between Lexicon-based and SVM has been implemented in several researches [5][3] where the opinion that used is written in English Language. In this research, the tweets that taken for the analysis are the tweets that is written using Indonesian Language which has a different structure with an English Language that is used on the previous researches.

II. LITERTURE REVIEW

A. Twitter

Twitter is a microblogging site that allow user to send their tweet with the maximum characters used are 140 characters. There are a lot of contents in Twitter such as, Profile, Following, Follower, Mentions (@), Direct Message, Hashtags, Trending Topics, etc [5]. In this research the data that is used to be analyzed is the tweet about television program that has been remove the URL, hashtag and also retweet (RT) that does not giving additional opinion in it.

978-1-4799-7752-9/15/$31.00 ©2015 IEEE

B. Television Program

Television program television program is a basic plan of a television show concept that will be the foundation of creativity and production design that will be divided into various main criteria that matched with the purpose and target audience of the show [6]. In Indonesia, the number of people who watch television programs play an important role in determining whether the television program will be extended or not. However the problem in Indonesia is the television program that has a high value of rating and share did not shown to have a high quality as well. This is why sentiment analysis focuses on the television program.

C. Lexicon-based method

Lexicon-based is belong to unsupervised techniques, this method classify the data into two classes of positive or negative [7] [3]. This lexicon-based method uses the help of dictionary to classify the tweet into positive sentiment or negative sentiment. There are some steps of lexicon-based that is used in this research, such as determining the polarity of words, negation handling, and also giving score to every each entity in the tweet.
The formula to calculate the score for the entity can be seen in equation (1).

$$score(e) = \sum_{wi:wi \in L \cap wi \in s} \frac{wi.so}{dis(wi,e)} \quad (1)$$

Explanation:
score(e)= The Final Label Score of the Entity
w_i = An Opinion Word
L = All Opinion Words
s =Sentence That Contains Entity and Opinion Words
so = The Label of The Opinion Word (+1 Or – 1)
dis(w_i,e)= Distance between Entity (e) and the Opinion Words (W_i)

D. Support Vector Machine

Support Vector Machine is belong to supervised techniques. Supervised Learning in sentiment analysis is a method that trains a sentiment classifier that is taken based on the frequency of occurrence of various words contained in the document, text, or tweet [8]. By doing training process that uses the data input in the form of numerical data such as word index number, and also the weight (usually obtained from the calculation of TF-IDF, Term Presence, etc.) will result in a value or pattern that will be used in the testing process for labeling process tweet. The purpose of Support Vector Machine method is actually to find the optimal hyperplane which has a maximum margin. Margin itself is the distance between the hyperplane (lines) with the closest point from each class, this closest point usually called as Support Vector [8]. The formula of hyperplane can be seen in equation (2)

$$f: w \bullet x + b = 0 \quad (2)$$

Explanation:
w = Hyperplane Parameters (the Perpendicular Line between the Hyperplane and the Point of Support Vector)
x = Data Input SVM (X_1 = Word Index, X_2= Word Weight)
b = Hyperplane Parameters (Bias)
f = Function of Hyperplane

Hyperplane parameter (w) obtained by using the formula in equation (3).

$$w = \sum_{i=1}^{l} \propto_i y_i x_i \quad (3)$$

Bias value (b) obtained by using the formula in equation (4).

$$b = \frac{b^{(1)} + b^{(2)}}{N + M} \quad (4)$$

The value of N, M, $b^{(1)}$, dan $b^{(2)}$ are the result from the formula in equation (5).

$$b^{(1)} = 1 - (\sum_{i=1}^{N} w_i \bullet x_i) \quad (5)$$

$$b^{(2)} = -1 - (\sum_{i=1}^{M} w_i \bullet x_i)$$

Explanation:
w = Hyperplane Parameters (The Perpendicular Line between the Hyperplane and the Point of Support Vector)
$\propto_i$ = Non-Negative Variable Lagrange Multiplier (Alpha)
y_i = Data Input (Label -1 Or +1)
x = Data Input SVM (X_1 = Word Index, X_2= Word Weight)
b = Hyperplane Parameters (Bias)
$b^{(1)}$ = Support Vector from Class +1
$b^{(2)}$ = Support Vector from Class -1
N = The amount of $b^{(1)}$
M = The amount of $b^{(2)}$

E. Evaluation of Sentiment Classification

Once the system has successfully clarify the tweet, we need a measure to determine how valid the classification that has been made by the system. Table I shows the confusion matrix which is used to assist in the calculation of the evaluation system [9].

Table I Confusion Matrix [9]

	Predicted positives	Predicted negatives
Actual positives instances	Number of True Positives instances (TP)	Number of False Negatives instances (FN)
Actual negatives instances	Number of False Positives instances (FP)	Number of True Negatives instances (TN)

The formula that is used as a reference indicator in assessing the system created can be seen in sub-section below [9]:

1. Accuracy
   The value of the entire true predicted against all predicted. The formula to obtain accuracy can be seen in equation (6).

$$Accuracy = \frac{TN + TP}{TN + TP + FP + FN} \quad (6)$$

2. Precision
   The value of true positive prediction against all positive prediction. The formula to obtain precision can be seen in equation (7).

$$\text{Precission} = \frac{\text{TP}}{\text{TP+FP}} \quad (7)$$

3. Recall
   The value of true positive prediction against all actual positive. The formula to obtain recall can be seen in equation (8).

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

III. METHODOLOGY

This stage will give more detailed explanation of the process that is done in this research. But before that, try to look at the general overview of the system in Figure 1.

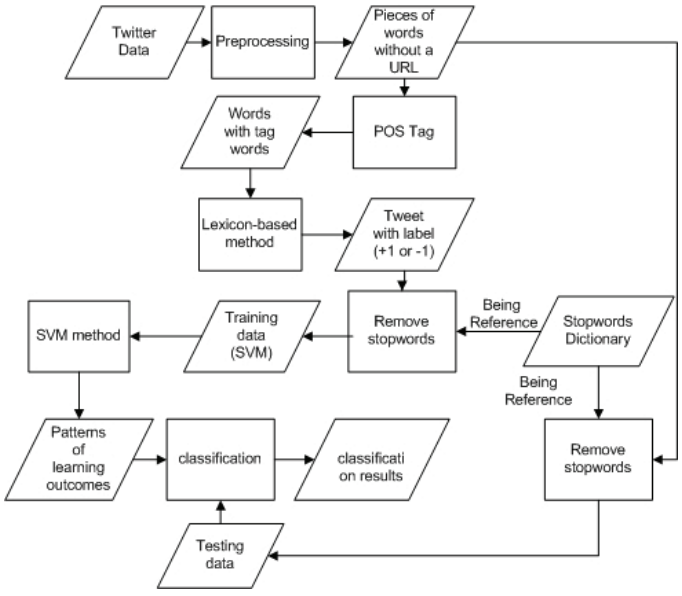

Fig. 1 General Overview of the System

Figure 1 is a picture of the process that is performed on the system. However, the steps will be explained more detail on the points in the next section.

A. Data Preprocessing

Twitter data that is taken usually still in a state of unstructured, that is why the unstructured data needs to be converted into data that can be processed by the preprocessing stage. The preprocessing that is used in this research: lowercase, tokenize, convert the slang words into the formal words, remove URL, and remove stopwords. Note that in this research, the process of removing stopwords is being processed after the lexicon-based method done and before implement the svm method.

B. Lexicon-based method

After the preprocessing is done, the data that has been cleaned then those data will be processed in lexicon-based method. Some of steps that used in this research can be seen in sub-section below [6]:

1. Extracting opinion words and giving the polarity value by matching words in tweet with words in the opinion words dictionary. When the word is found in the opinion words dictionary, then the label (+1 or -1) that exist in dictionary will be used as the value of the polarity for the opinion word.

2. Handle negation by searching for words that indicates negations such as “kurang”, “tidak”, etc. When the negation word is located before the opinion word, the value of the polarity of opinion words will be replaced with the value of the opposite of the polarity before. So if the value of the previous polarity is +1, then the value will be changed to -1, and the reverse.

3. Calculate the distance between entities opinion words by identifying the location of the entity (usually a noun), and opinion word (verbs, adjectives, adverbs, and nouns). After that calculate the distance between the entity and opinion word using equation (1).

When all three processes performed successfully, the output will be a tweet that has been in classification or a tweet that has been labelled. This tweet is the one that used as training data in the process of SVM.

C. Support Vector Machine

In this process, the result from lexicon-based method is being used as the training data, while the testing data is taken from the result of preprocessing. Before the data is further processed, the data must go through the step of removing stopwords so only the important words are processed in this method. Svm method requires a weighting process to change the word into a number so that the word can be processed in a vector space. The weighting words that used in this research in weighting term frequency – invers document frequency (TF-Idf), with feature-based unigram.

The steps svm conducted in this research are as follows:

1. Prepare the input data in the form of index word, the weight of word, and also its label
2. Calculate the parameter w by using the equation (7)
3. Calculate the bias (b) by using the equation (4)

4. Getting the classification for data testing using the equation (2)

Those are the four steps of the process that is performed on svm method in this research.

IV. TESTING

When all stages successfully performed and the system is able to pull out the result, then the next thing to do is to evaluate the results obtained from the system. In this evaluation, assessment can be seen from the values of accuracy, precision and recall

A. Dataset

This study uses five television programs as the example. The tweet for each television programs is selected randomly, and then the data is divided into training data and testing data. Table II shows the dataset used in this research.

Table II Datasets

Television Program	Total Tweets
Program A	300
Program B	300
Program C	500
Program D	500
Program E	600

B. Testing Models

1) The Scenario testing on comparisons of training data and testing data that are used. The details of the comparison of both these data are as follows:
   1. Training data 60%: Testing data 40 %
   2. Training data 70%: Testing data 30 %
   3. Training data 80%: Testing data 20 %
   4. Training data 90%: Testing data 10 %

2) The Scenario testing of the conclusions obtained from the data tweet that has successfully analyzed. This testing model will shows the conclusion on the amount of positives and negatives sentiments obtained from each television program. The conclusion is only taken from the testing data that get the highest accuracy in each television program.

C. Testing Results

The testing is done manually, by matching the label from the results of the system with the label that has been given manually. To avoid subjective assessment, this research is assisted by five people who have more expertise in Indonesian language. Table III shows the results of accuracy obtained for each television program and also the comparisons ratio of training data and testing data that are used. By looking on the result of the comparisons of training data and testing data, it shows that those comparisons do not affect the value of accuracy. This is proved by the value of accuracy that tends to have an unstable increase and decrease value.

However, when the average value of accuracy is calculated in each comparisons, it shows that the highest average values for accuracy falls on comparison of training data 90%: testing data 10%.

Table III Accuracy value

Television Program	Accuracy			
	Training 60%: Testing 40%	Training 70%: Testing 30%	Training 80%: Testing 20%	Training 90%: Testing 10%
Program A	76.67	71.11	74.17	70.00
Program B	71.25	72.22	70.00	80.00
Program C	61.25	61.67	67.50	62.00
Program D	53.75	56.33	59.00	58.00
Program E	50.21	54.17	49.58	52.50

Table III shows that the highest accuracy occurs in Program A and Program B, while Program C and Program D tend to have a low value in accuracy. Based on the result of sentiment analysis on testing data, Program A and B are televisions program which dominated by the positive opinion, while Program C, Program D, and Program E are programs that dominated by the negatives opinions according on the testing result. So, according to the analysis result it can be conclude that television programs that have more positive opinions tend to have a better accuracy than the one that have more negative opinions.

The precision and recall analysis value can be seen on Table IV. The Precision and recall values referenced in this research are the value of precision and recall the positive class, the equation from this evaluation value can be seen on equation (7) and equation (8).

According to Table IV, the precision value for Program A and Program B (programs that dominated by positive opinions) is on range of 70% to 80%. While Program C, Program D, and Program E (programs that dominated by negative opinions) have the precision value on range 49% to 88%.

There are two reasons why the high precision value can be obtained. First, the system can correctly classify positive tweets. Second, the system performs a small error in the classification of negative tweets. This analysis can only be determined by looking at the recall value.

From Table IV it can be seen that the recall value for Program A and Program B (programs that dominated by positive opinions) is on range of 97% to 100%. While Program C, Program D, and Program E (programs that dominated by negative opinions) have the precision value on range 7% to 33%.

If analysis on the value of precision and recall values are combined, it can be concluded that :

1. Program A and Program B (programs that dominated by positive opinions) which have a high values on precision and recall, so it means the system can correctly classify positive tweets, and tends to have small error in the classification of negative tweets.

2. Program C, Program D, and Program E (programs that dominated by negative opinions) which have high values on precision but low value on recall, it means that the system performs a small error in the classification of negative tweets, but the system cannot classify the positives tweet correctly, it can be seen from the low recall value obtained.

Table IV Accuracy, Precision, Recall

Television Program	Training 60%:Testing 40%			Training 70%:Testing 30%			Training 80%:Testing 20%			Training 90%:Testing 10%		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Program A	76.67	76.79	99.44	71.11	72. 33	97.73	74.17	73.94	100.00	70.00	70.00	100.00
Program B	71.25	71.25	100.00	72.22	73.04	98.51	70.00	70.00	100.00	80.00	80.00	100.00
Program C	61.25	49.85	11.59	61.67	75.00	13.53	67.50	83. 33	18.50	62.00	75.00	7.63
Program D	53.75	63. 33	11.78	56. 38	81. 32	17.88	59.00	81.85	23.33	58.00	88.46	20.69
Program E	50.21	63.41	22.66	54.17	67. 38	27.12	49.58	75.69	18.21	52.50	60.00	33.71

V. CONCLUSIONS

After all the process has been done, it can be concluded that in this research the television program that dominated by positive sentiments is the one that has the highest accuracy, which is 80%, whereas the program that consist negative sentiments have a low accuracy. Although the highest precision value (with 88.46%) found on Program C (training 90% : Testing 10%), the accuracy on that Program still on a low value. This happened because the recall value obtained on Program C is low. So the high value of precision obtained because the system performs a small error in the classification of negative tweets, but the system cannot classify the positives tweet correctly, that is why the accuracy and recall value are low.

Moreover, the results show that the ratio of training data and testing data that has been used is not influenced the value of the accuracy, precision, and recall. It is proved by the result that shows the increase and decrease value on each ratio did not change linearly.

Based on the result, the combination of these two methods has not giving a good result, therefore a lot of things that can be done to improve the quality of this method. The lexicon-based method can be improved by using a dictionary that can provide labels according to the level of strength from each opinion word that represent positive or negative sentiment, as contained in the WordNet or SentiWordNet dictionary. This is suggested because Indonesian structure are slightly difficult to analyzed by using the basic/usual labeling (using +1 or -1 labeling). As for the Support Vector Machine (SVM) method, it can be improved by using other parameters such as additional slack parameters, or changing the dimensions into the higher dimensions by using other kernels such as polynomial or RBF.

ACKNOWLEDGMENT

The author would like to thank those who have helped, especially for Alfan Farizki and Ayu Purwarianti who has been given permission to use the dictionary that has been made in their research.

References

[1] Yayasan, SET., Yayasan, TIFA., IJTI. The Habibie Center & LSPR. "Menuju Televisi yang Ramah Keluarga". Rating Publik

[2] Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

[3] Tan,S., Wang, Y., & Cheng,X.(2008). "Combining learn-based and lexicon-based techniques for sentiment detection without using labeled examples", In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, July 20-24, 2008, Singapore, Singapore

[4] Pang, B., Lee, L., & Vithyanathan, S. (2002). "Thumbs Up? SentimentClassification Using Machine Learning Techniques." *Proceedings of The ACL-02 conference on Empirical methods in natural language processing* (pp. 79-86). Stroudsburg: Association for ComputationalLinguistic.

[5] Ley, Z., Riddhiman, G., Mohamed, D., Meichun, H., & Bing, L. (2011). "Combining lexicon-based and learning-based methods for twitter sentiment analysis". *HP Laboratories*, Technical Report HPL-2011, 89.

[6] Rukmananda, N. (2004). *Menjadi sutradara televisi: dengan single dan multi camera.* Grasindo, Indonesia.

[7] Ding, X., Liu, B. & Yu., P. S. (2008), "A Holistic Lexicon-Based Approach to Opinion Mining". In Proceedings of First ACM International Conference on Web Search and Data Mining (WSDM-2008)

[8] Croft, B., Metzler, D., Strohman, T. (2009). *Search Engines: Information Retrieval in Practice*, Addison-Wesley Publishing Company

[9] Khairnar, J., Kinikar, M. (2013). "Machine Learning Algorithms for Opinion Mining and Sentiment Classification". International Journal of Scientific and Research Publications, 3 (6), 1-6.