**TESIS**

**OPINION MINING PADA TWITTER UNTUK BAHASA INDONESIA DENGAN METODE SUPPORT VECTOR MACHINE DAN METODE BERBASIS LEXICON**

**OPINION MINING ON TWITTER FOR INDONESIAN USING SUPPORT VECTOR MACHINE AND LEXICON BASED METHOD**

****

**JAN KRISTANTO WIBISONO**

**10/306141/PPA/03230**

**PROGRAM STUDI S2 ILMU KOMPUTER**

**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM**

**UNIVERSITAS GADJAH MADA**

**YOGYAKARTA**

**2013**

# TABLE OF CONTENTs

[DAFTAR ISI v](#_Toc367271047)

[BAB I 1](#_Toc367271048)

[PENDAHULUAN 1](#_Toc367271049)

[BAB II 3](#_Toc367271050)

[TINJAUAN PUSTAKA 3](#_Toc367271051)

[BAB III 6](#_Toc367271052)

[METODOLOGI PENELITIAN 6](#_Toc367271053)

[3.1 Pengumpulan Data 6](#_Toc367271054)

[3.2 Analisis Sistem 6](#_Toc367271055)

[3.3 Garis Besar Sistem 8](#_Toc367271056)

[BAB IV 10](#_Toc367271057)

[HASIL PENELITIAN DAN PEMBAHASAN 10](#_Toc367271058)

[4.1 Perbandingan proses pelabelan data pelatihan 10](#_Toc367271059)

[4.2 Akurasi Pemberian label 12](#_Toc367271060)

[BAB V 13](#_Toc367271061)

[KESIMPULAN DAN SARAN 13](#_Toc367271062)

[5.1 Kesimpulan 13](#_Toc367271063)

[5.2 Saran 13](#_Toc367271064)

[DAFTAR PUSTAKA 15](#_Toc367271065)

# CHAPTER I

# INTRODUCTION

Opinion is a personal opinion that is not objective and does not go through the verification process [1]. Other people's opinions about a matter to be important in decision making. Suppose someone would buy a car normally he would ask other people what they thought about that car. Or often when we buy certain products, we are welcome to fill in the questionnaire related to the product. Nowdays does not need to directly ask about the opinions of others directly. Along with the development of the internet and the web, a lot of people had shared information about their opinions and experiences even their critique of a particular product.

Social media has grown rapidly, starting from the beginning era of Friendster to Facebook, Twitter, Google+ and Foursquare. According socialbakers.com Facebook users in Indonesia has reached 42.5 million, making Indonesia a Facebook user ranked fourth after the United States, India and Brazil. Twitter users in Indonesia is also not less in number that reached 19.5 million [2] which ranks fifth after the U.S., Brazil, Japan and the UK. This rapid development make people communicate easily. They shared information, sharing about their daily activities or to provide notice to their community, and even their grievances towards something.

Once the popularity of social media, social media users now not only the personal or individual. But the corporate also use social media as a medium of communication with the people. Such as the introduction of new products, product offerings, and even the legislature also promote themselves through social media. Cheaper than those considered in terms of cost, and also right on target.

Based on the above there is a huge potential exposure to social media to do opinion extraction or called opinion mining, to find out what others think of a thing, which can be used during the decision making process. Several studies already exist today mostly uses both machine learning with naïve Bayes [3], [4], [5], [6], [7], Support Vector Machine [4], [5], [8], [6] and Maximum Entropy [6] or lexicon-based [9]. This research will be conducted on opinion mining by combining machine learning-based approach and based lexicon. Methods that will be used to represent the machine learning approach is Support Vector Machine (SVM), which according to the paper Zhang et al. [8] often provide the best accuracy compared with other learning methods. With a combination of methods of Support Vector Machine (SVM) based lexicon and is expected to improve the accuracy of the classification process of opinion.

# CHAPTER II

# LITERATURE REVIEW

This chapter contains a description of the studies that have been done previously with regard to opinion mining or sentiment analysis in English and Indonesian.

Pang et al. [4] uses machine learning to classify movie reviews. This study determined the classification of the sentiments of the movie review and determine whether the review has a positive sentiment or negarif. Different features extracted from the movie reviews and used machine learning algorithms Naïve Bayes and Support Vector Machine (SVM) to generate classification models. They earn between 78.7% accuracy when using a Naïve Bayes on the use of unigram. Accuracy obtained when using SVM with unigram is 72.8%.

Franky and Manurutng [5] tried to repeat the experiment sentiment classification of movie reviews done by Pang et al. [4] The main difference of this study, this study takes the focus adalalah movie reviews in Indonesian. Related to the absence of a corpora traning for Indonesian, it was applied to machine translation tools to translate the English corpus made ​​Pang et al. [4] which is native to the Indonesian and translasinya results are used, ranging from commercial tool to simple word-for-word translation and text classification methods try. Average yield obtained for the Naïve Bayes method was 74.62% and 75.62% for the SVM method.

Opinion mining in the language of china was made by Zhang et al. [8]. This study, also inspired from research Pang et al. [4]. If the research Pang et al. [4] using the data movie reviews in English, this study used computer online product reviews in Chinese language. They developed the opinion mining based learning methods. Using an existing dataset on AmazonCN opinion. The learning method used is a Support Vector Machine using a string kernel.

In 2009 when social media began to demand internet users. Go et al. [6] perform opinion mining in social media twitter. This study start up idea makes use of emoticons in labeling training data from English-language tweets. This idea complements a previous study conducted Pang et al. [4] who still use manual way to get the training data. This study classifies the tweet sentiment on 2 classes, namely positive and negative sentiment classes. Accuracy obtained is 81.3% by using Naïve Bayes and 80.5 using Maximum Entropy and 82.2% using SVM.

Pak and Paroubek [7] using emoticons to build English language training data from twitter with positive sentiment, negative and neutral. The idea of ​​building a training data using emoticons actually been instigated by Go et al. [6]. This study adds to the neutral class training data take account of data from your English-language media. The method used is Naive Bayes with n-grams. Best Performasi generated when using bigram.

Komarsilam and Winarko [9] perform opinion mining in Indonesian language movie reviews by unsupervised learning method based on Indonesian-language lexicon and use the rule to determine the opinion phrases and entities relating to such opinion. Results obtained in the form of a text sentiment polarity is positive, negative or neutral opinion and accompanying entities specified, where the analysis is described in a diagram that shows the number and percentage of sentiment labels also display the detail of the results of sentiment analysis. From the test results of the system obtained precision values ​​for 0616 and 0643 amounted recall value.

Alliandu and Winarko [3] perform on twitter Opinion mining in Indonesian language. The method used is the naïve Bayes supervised learning. Methods of data collection takes the idea of training and research Pak and Paroubek [7] that using emoticons and method for retrieving data from the class of neutral media accounts. Feature weighting used is the laplace smoothing term frequency and TF-S. The results were obtained accuracy of 77.45% for the laplace smoothing term frequency and accuracy of 75.86% for TF-IDF in the test set were annotated using emoticons.

# CHAPTER III

# RESEARCH METHOD

## Data Collection

Data required in this study consists of two types of data, ie data used for training and testing data labeling. The data used for training in the form of documents tweet Indonesian language that is labeled with the lexicon-based method. This document was taken from research Alliandu and Winarko [3] that the number 81500 tweet. However, after being processed using lexicon-based data used for training only 25000.

Unlike the data used for training, the data used in the labeling process using tweets that do not have labels. Tweet document is obtained by utilizing the twitter API can search for existing tweets on twitter with a certain parameters.

## Analysis System

Opinion Mining System has two main processes , namely training and grading sentiment on the new data . Training is used to get the model to be used for classroom menentuan sentiment on the new data . Processes that occur in these systems is shown in Figure 1.

In the training process by using the feedback form that is used in the research corpus Komarsilam and Winarko [ 9 ] has been done dengen labeling method based lexicon . The initial step of labeling by using lexicon -based preprocessing is done on the data . In preprocessing , there are several stages , in order are cleaning karater not including UTF - 8 , change all characters to lowercase , remove the symbols , process numbers , remove the loop character , the replacement of the word is not raw , do spelling correction . After completion of the preprocessing stage , the next resistance will be included in the determination of the label . In this stage there are also several processes including POS taging , analyze tweet with existing rules , negation check . Up here that the label has been generated corpus is determined by using lexicon -based and ready to be used for the training process .

The next process is the core process of training , in this study there were 25,000 training data with the label prescribed lexicon -based method , the data consists of 12,500 positive and 12,500 labeled data is data is labeled negatively. The training process will produce SVM models . That is a model that will be used to determine the sentiment of the SVM method.

The second main process is the determination of a new tweet with the label SVM method . The first step in the main process is the system will search for tweets related to desired keywords . Next will be entered in a preprocessing stage which has been described previously. Then checking the presence of a phrase in your opinion who has dipreprocessing . Furthermore, to tweet a phrase that indicated to have opinions , sentiments will be determined by the method of SVM .

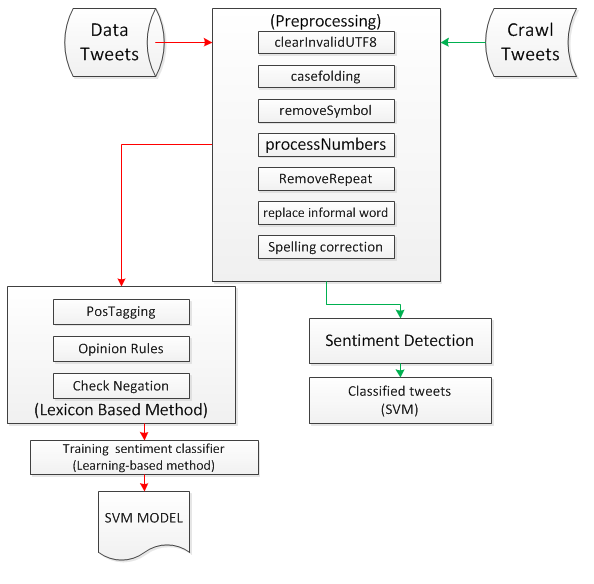


Figure . Process Opinion Mining

## Outline of The System

Opinion Mining system architecture that is made is shown in Figure 2 At the system that created Opinion Mining are the two major processes. Ie the training process and labeling process of determining or testing. Training process using lexicon-based method for determining the label, which then results will be labeling the training data in the testing process that uses Support Vector Machine algorithm.

Groove on the training process is done proprocessing on existing data. Where in the corpus preprocessing involves formal and informal words, which then would produce data that tweet was in terms of data preprocessing or the net. Upon completion of the preprocessing will be performed to determine the label on the data that has been dipreprocessing tweet using lexicon-based methods. Up here already have a data system that already has a label tweets. Furthermore, the data that your own labels will be converted into vectors as training data SVM using TF-IDF method. Then do the training process with SVM methods. This will result in SVM training model that will be used in the process of determining the label or testing.

Workflow process is the process of determining the next big label or testing. The first system will search for tweets using tweetAPI that match the keywords the user entered. This process will result in a tweet and the data will be stored in the database. Then performed on the data preprocessing the tweet. Prerpocessing done the same as it did in the training process flow. After the preprocessing will be checked phrase, whether or not it contains opinion phrases. If the opinion contains a phrase that will be included in the classification process.

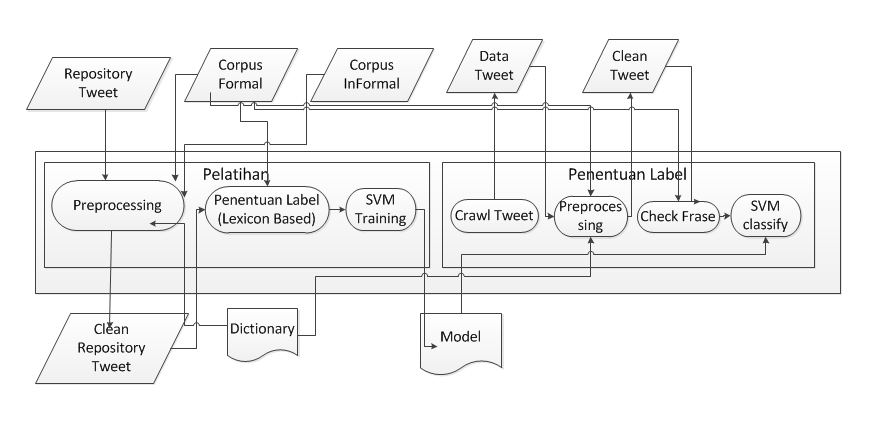


Figure . System Architecture

# CHAPTER IV

# RESULTS AND DISCUSSION

Results of the study consisted of three parts. First, the comparison between the training data labeling process is done Alliandu and Winarko [3] that using emoticons and labeling methods of training data is done by the method based lexicon. Second, labeling experiments using a combination of lexicon-based methods and support vector machine. Third, the comparison of labeling and labeling methods combined with manual.

## Comparison of training data labeling process

Data labeling process conducted in this study will use a lexicon-based method of data while taking training from the training conducted by Alliandu and Winarko [3]. On training conducted Alliandu and Winarko [1] method is the method used emoticons. Emoticon method is to detect certain emoticons, emoticon that each has a correlation with the label.

The experiments were performed using 300 random data taken previously been labeled on Alliandu and Winarko study [3]. In this experiment that compared the accuracy of the method is used emoticons on Alliandu research and Winarko [3] and compared to manual labeling lexicon-based methods performed in this study and manual labeling. In the tests performed on the emoticon and labeling method has an accuracy of 0:49 manuals that detail shown in Table 1.

Table 1. Comparison of Emoticon Methods and Manual

|  | **Manual** | | | |
| --- | --- | --- | --- | --- |
| **Emoticon** |  | **Positive** | **Neutral** | **Negative** |
| **Positive** | 40 | 46 | 14 |
| **Neutral** | 12 | 57 | 31 |
| **Negative** | 11 | 40 | 49 |
|  | **63** | **143** | **94** |

In the tests performed on the lexicon-based methods and has an accuracy of 0.68 manual labeling are shown in detail in Table 2.

Table 2. Comparison of Lexicon Methods and Manual

|  | **Manual** | | | |
| --- | --- | --- | --- | --- |
| **Lexicon** |  | **Positive** | **Neutral** | **Negative** |
| **Positive** | 36 | 13 | 8 |
| **Neutral** | 23 | 119 | 37 |
| **Negative** | 4 | 11 | 49 |
|  | **63** | **143** | **94** |

Based on two experiments that have been conducted, labeling is done with the lexicon-based method proved to have better accuracy is 0.68 compared labeling is done by the method of emoticons that is only 0:49.

## Accuracy of labeling

Calculation accuracy of this study was to compare the results of the labeling is done with a combination of lexicon-based method and the SVM compared with the labeling is done manually. The data used for comparison is data from 3 trials with input keywords different. Based on experiments conducted 3 499 Data obtained tweet. Table 33 shows the results of the comparison has been done. 6.9 In the picture shown of 499 tweets tweet number labeled by the system is properly numbered 388 tweets. Based on these data, the accuracy of the study was 0.78

Table 3. System accuracy

|  | **Manual** | | | |
| --- | --- | --- | --- | --- |
| **Emoticon** |  | **Positive** | **Neutral** | **Negative** |
| **Positive** | 40 | 46 | 14 |
| **Neutral** | 12 | 57 | 31 |
| **Negative** | 11 | 40 | 49 |
|  | **63** | **143** | **94** |

# CHAPTER V

# CONCLUSION, SUGGEST AND REFERENCE

## Conclusion

Based on test results that have been obtained, it can be deduced:

1. In the process of collecting training data lexicon based method provides better accuracy than the methods emoticons. In this study methods based lexicon reaches 68% accuracy, while the method of emoticons just reached 49% accuracy. This test is based on 300 random tweets Indonesian language contained in all tweet emoticons.

2. At the end of testing a combination of lexicon-based methods and support vector machine produces 78% accuracy. This test is based on 499 random tweets Indonesian language taken directly from twitter.

## Suggest

In this study postagging used method is to use a dictionary. Turns postagging results using less good dictionary. Though used in a lexicon based postagging which is the method used for data collection training. It required that postagging method has higher accuracy. So the results of the accuracy of data collection can be better again.

Lexicon method based on phrase matching with sentiment data dictionary is that of the single. It means the word has only one sentiment. Though it is possible one word can have more than one sentiment. That requires a dictionary word each sentiments he may have more than one sentiment.

This research uses data on training as much as 25000. It turns out the training data is not enough. The addition of training data can improve the accuracy of the classification is done by SVM

[1]Quirk, R., Greenbaum, S., Leech, G. dan Svartvik, J., 1985. A Comprehensive Grammar of the English Language, Longman, London..

[2]Semiocast, 2012. Brazil becomes 2nd country on Twitter, Japan 3rd Netherlands most active country. http://semiocast.com/en/publications/2012\_01\_31\_Brazil\_becomes\_2nd\_country\_on\_Twitter\_superseds\_Japan, diakses 30 Mei 2012.

[3]Alliandu, P. dan Winarko, E., 2012. Analisis Sentimen Tweet Berbahasa Indonesia di Twitter. Tesis. : Program Pasca Sarjana Universitas Gadjah Mada.Yogyakarta.

[4]Pang, B., Lee, L. dan Vaithyanathan, S., 2002. Thumbs up?: Sentiment Classification Using Machine Learning Techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10. Stroudsburg.

[5]Franky dan Manurung, R., 2008. Machine Learning-based Sentiment Analysis of Automatic Indonesian Translations of English Movie Reviews. In Proceedings of the International Conference on Advanced Computational Intelligence and Its Applications.

[6]Go, A., Bhayani,R. dan Huang, L., 2009. Twitter Sentiment Classification using Distant Supervision, CS224N Project Report, Standford

[7]Pak, A. dan Paroubek, P., 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. European Language Resources Association, Valletta.

[8]Zhang, C., Zuo, W., Peng, T. dan He, F., 2008. Sentiment Classification for Chinese Reviews Using Machine Learning Methods Based on String Kernel. In Proceedings of the 2008 Third International Conference on Convergence and Hybrid Information Technology. Washington, DC.

[9]Komarsilam, E. dan Winarko, E., 2012. Penambangan Opini Pada Situs Review Film Berbahasa Indonesia, Tesis, Program Pasca Sarjana FMIPA UGM, Yogyakarta.