A Framework for Sentiment Analysis Implementation of Indonesian Language Tweet on Twitter

Asniar[1] and B R Aditya[2]

[1,2] School of Applied Science, Telkom University, Jl. Telekomunikasi No.1, Bandung, Indonesia

Email: [1]asniar@telkomuniversity.ac.id, [2] bayu@tass.telkomuniversity.ac.id

Abstract. Sentiment analysis is the process of understanding, extracting, and processing the textual data automatically to obtain information. Sentiment analysis can be used to see opinion on an issue and identify a response to something. Millions of digital data are still not used to be able to provide any information that has usefulness, especially for government. Sentiment analysis in government is used to monitor the work programs of the government such as the Government of Bandung City through social media data. The analysis can be used quickly as a tool to see the public response to the work programs, so the next strategic steps can be taken. This paper adopts Support Vector Machine as a supervised algorithm for sentiment analysis. It presents a framework for sentiment analysis implementation of Indonesian language tweet on twitter for Work Programs of Government of Bandung City. The results of this paper can be a reference for decision making in local government.

1. Introduction

In the digital era, people have a variety of social media accounts such as Facebook, Twitter, Instagram, Path, Pinterest, Blogs, and other social media. Every time these people put the idea to share on social media or leave a comment. With the social media, people as if finding media to express opinions more easily and more freely. Besides that, people prefer to express what they think in social media rather than express directly to the object or the person.

Along with the widespread use of social media, the data become more and more unsystematic. Sometimes the writing is shared socially mediated consists of only one word that is difficult to understand or not rare that someone makes writing that does not mean anything. Millions of digital data are scattered and are still not used to be able to provide any information that has value to the government or the company.

Sentiment analysis that is part of the opinion mining [4], is the process of understanding, extracting and processing the textual data automatically to get information [2]. Conducted to see opinion on an issue, or it can also be used to identify the tendency of things in the market [1]. Sentiment analysis in this research is the process of textual document classification into two classes, positive and negative sentiment. The magnitude of the effect and benefits of sentiment analysis, leading research or applications on the rapidly growing sentiment analysis, even in America, approximately 20-30 companies that focus on sentiment analysis service [4]. Basically, sentiment analysis is a classification, but the reality is not as easy as usual because the classification process is related to the use of language. In which there is ambiguity in the use of words, the absence of intonation in a text, and the development of language itself [3].

The interaction between leaders and members is a necessity in order to create synergy and can produce a business or product that fits the vision and mission. For example in this case is the interaction between the Municipal Government, specifically in this case the mayor of officials and citizens. There are many ways to interact with the mayor, through a forum, official letters, letters of complaint, demonstrations, Facebook pages, and even Twitter. Interaction with twitter carried out by one mayor in Indonesia, namely Mr. Ridwan Kamil who is the mayor of Bandung City. Mr. Ridwan Kamil almost every day using Twitter as a means to interact with citizens. His twitter account is @ridwankamil. According to the observations, Mr Ridwan Kamil receives a response from the public ranging from around dawn to midnight. Tweet from Mr. Ridwan Kamil usually contains about socialization of program of the government (local government) and response about complaints from the public.

Sentiment analysis can be performed to monitor Bandung City Government's Programs which quickly can be used as a tool to see the public response to that program so that the next strategic steps can be conducted. Based on this background, it is critical to conduct research about sentiment analysis towards Indonesian language documents of Bandung City Government's Programs.

Researches about sentiment analysis of tweet have been conducted by many researchers such as "Twitter as a Corpus for Sentiment Analysis and Opinion Mining" [8] which provides a method of data collection. Furthermore, the research entitled "Analisis Sentimen Pada Dokumen Berbahasa Indonesia dengan Pendekatan Support Vector Machine" [6] is added language detection process to get data in Indonesian language and to produce result that the classification process that is supervised by an algorithm using Support Vector Machine (SVM) has level of accuracy that is better than the Naïve Bayes. Unfortunately, no study has tried to make the implementation of these two studies in the form of software tools that can automatically perform sentiment analysis of Indonesian tweet to see public opinion on Bandung City Government's Program. However, twitter is already used by Mayor of Bandung City to interact with the community. So it is very important software tools that can rapidly process and analyze data in social media tweet related to Bandung City Government's Programs. By doing so, Mayor of Bandung City can make policie regarding to its work program quickly.

Through this study, researcher can propose a framework of implementation of sentiment analysis of Indonesian language tweet on twitter for Work Programs of Bandung Government so that it can become a reference in the creation of applications sentiment analysis for Indonesian language tweet for work program of Bandung City.

In summary, the contribution of this work is described as follows Opinion Lexicon, Sentiment Analysis, and Proposed Framework for Implementation of Sentiment Analyisis of Indonesian Language Tweet on Twitter.

2. Opinion Lexicon

In the process of sentiment analysis, document is identified whether it is an opinion (sentiment contained in the document). To get the sentiment, the document should be analyzed so that it can be concluded whether it contains sentiment or not. Opinion lexicon is a collection of words containing sentiment. Examples of words containing opinions with positive sentiment are "cantik" , "indah", "bagus" and "luar biasa". While examples of words that contain the negative sentiment are "jelek", "lemah" and "payah" [4].

In practice, the use of word having opinion consists of two kinds [2] :

2.1. Direct Opinion

Direct opinion is a direct expression of the sentiments of the target object such as product, topic or person in a document.

2.2. Comparative Opinion

Comparative opinion is to compare the similarity or difference of more than one object.

The technique of collecting sentiment words can use dictionary-based approach [4]. The strategy used is listing words that are previously known their sentiment, then conducting a query to the dictionary to get the equation of words (synonyms) or opposite (anonymous). Query results are then used as input parameters for the next query, until no more new word again. For English dictionary, the commonly used dictionary is WordNet, which has already contained full lexical English. The weakness of this technique is not able to get a word in accordance with a particular domain. To eliminate it, then a technique based on corpus is used in which word identification is done on a corpus with certain existing domain. The problem that arises is there is no corpus with a complete vocabulary for all languages.

3. Sentiment Analysis

Sentiment analysis is the process of understanding, extracting and processing the textual data automatically to get the sentiment information contained in an opinion sentence [2]. Sentiment analysis is one part of the opinion mining [4]. Sentiment analysis is conducted to see opinions towards a problem or an object by a person to view negative or positive [3]. One example of the use of sentiment analysis in the real world is the identification of market trend and market opinion to an object.

The magnitude of the effect and benefits of sentiment analysis led to the rapid growth of research and applications of sentiment analysis. Even in America, there are about 20-30 companies that focus on sentiment analysis service [4]. Pang et al in 2002 classified the review of the film on the level of a document that has positive or negative opinion by using supervised learning techniques. A set of movie reviews that had previously been determined to be either positive or negative is used as training data for a machine learning algorithm that is already there. Accuracy obtained ranged from 72% to 83% [1].

Outline techniques for performing sentiment analysis is divided into two [5]:

3.1. Using Engineering Symbols

In this technique, analysis of every word in the document and extracting relationships to get the sentiment are conducted. To use this technique, the meaning of each word must be known and the rules of words in a sentence varies depending on the language used. Examples of the use of symbols techniques such as by using a dictionary as a base, where a document is declared as a representation of the word. Each word is determined its sentiments then the use of certain functions (sum, average, weight etc.) is used to get the sentiment of a document. The use of this technique is highly dependent on the language used, because the rules of words is different in languages. Even though the standard rule has been formulated in language, in fact, it cannot be applied completely. It is associated with the language itself that evolves.

3.2. Using Machine Learning Techniques

In this technique the features used are unigram and n-gram [8]. In unigram features, words and symbols in the document are represented into vector shapes, and each word or symbol is counted as a single feature. Whereas the n-gram features use syntactic, semantic, link-based and part of speech. In this case the features taken are not word by word but by whole document, in which the relationship of each word in a document is analyzed previously to obtain the relation among words. The words that make up relation are taken into a feature classification. Machine learning technique emphasizes the use of statistic to process the texts. The examples of the use of machine learning techniques are the use of Super Vector Machine (SVM), Naïve Bayes, Maximum Entropy, Centroid Classifier and K-nearest Neighbor.

4. Implementation of Sentiment Analysis of Indonesian Language Tweet on Twitter

To implement sentiment analysis of Indonesian language tweet on twitter for work programs of Bandung City Government, it is necessary to develop the device application with the proposed framework includes the following steps as shown in figure 1.

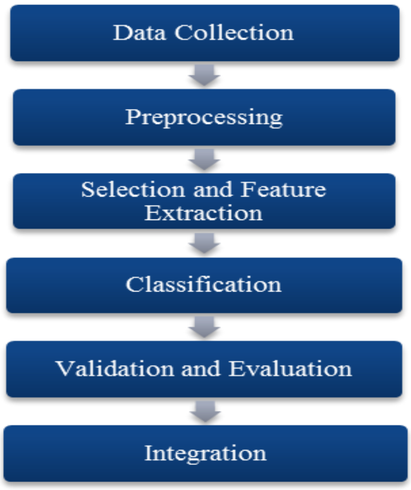

Figure 1. Proposed Framework for Sentiment Analysis Implementation of Indonesian Language Tweet on Twitter for Work Programs of Bandung City Government.

4.1. Data Collection

At this stage, software implementation will be made for data collection. The data used come from Twitter, a social networking site. Topics are limited regarding the comments about the Government's Work Programs in the City of Bandung in era of Mr. Ridwan Kamil. Methods of data collection used refers to the study “Twitter as a Corpus for Sentiment Analysis and Opinion Mining” [8] and “Analisis Sentimen Pada Dokumen Berbahasa Indonesia dengan Pendekatan Support Vector Machine”. [6] For the process of crawling, it uses the keyword of several Special Work Program, such as: the city park, culinary night, etc.

4.2. Preprocessing

At this stage, software implementation for preprocessing will be made. The purpose of preprocessing is to change the unstructured data into structured data.

The Methods that will be implemented at this stage are as follows [6]:

4.2.1. Cleansing. Cleansing is the process of cleaning of the document that contains words that are not required to reduce noise. The word omitted is HTML code, keyword, emotion code, hashtag (#), username (@ username), url (http://situs.com), and email (nama@situs.com).

4.2.2. Case folding. Case folding is the uniformity of letter as well as the elimination of the numbers and punctuation. In this case, only latin letters between a through z are used.

4.2.3. Parsing. Parsing is the process of breaking document into a word. This is according to the features used, namely unigram.

4.3. Selection and Feature Extraction

At this stage, software implementation will be made for the selection and feature extraction. The purpose of the selection and feature extraction is to create and classify data set which contains sentiment words.

The Methods that will be implemented at this stage are as follows [6]:

4.3.1. Part of Speech (POS). Part of Speech (POS) tagger is the process of giving a class to the word. Class of word selected is the word that contains many sentiments. Determination of the class of words is based on Kamus Besar Bahasa Indonesia (KBBI).

4.3.2. Stemming. Stemming aims to reduce variations of words that have the same basic word. As in the POS Tagger processing, stemming processing is also done based on KBBI.

4.4. Classification

At this stage, the implementation of the software will be made to do the weighting and classification. The purpose of weighting and classification is to determine the number of frequency of occurrences of positive sentiment and negative sentiment. The weighting processing is carried out by using Feature Term Frequency (TF), Feature Term Presence (TP), and the Term Frequency-Inverse Document Frequency (TF IDF). The classification processing uses supervised method with Support Vector Machine (SVM) algorithm because it has level of accuracy that is better than the Naïve Bayes according to the result of research “Analisis Sentimen Pada Dokumen Berbahasa Indonesia Dengan Pendekatan Support Vector Machine” [6].

Basically, SVM method classify data into two classes by constructing the N-dimensional hyper plane [10]. SVM uses g(x) as the discriminate function where the formula of g(x) is as follows: [11],

$$g(x) = w^T f(x) + b \tag{1}$$

where w is the weights vector, b is bias and f(x) is nonlinear mapping from input space to hogh-dimensional feature space. The parameters w and b are learned automatically on the training dataset following the principle of maximized margin using the following formula :

$$Min\ ½\ w^T w + C \sum_{i=1}^{N} ci \tag{2}$$

where N is the slack variables and C is the penalty coefficient.

4.5. Validation and Evaluation

At this stage the software implementation will be created to perform the validation and evaluation by calculating the accuracy with coincidence matrikx showing by Table 1[12].

Table 1. Coincidence Matrix [12]

		True Class	
		Positive	Negative
Predicted Class	Positive	True Positive Count (TP)	False Positive Count (FP)
	Negative	False Negative Count (FN)	True Negative Count (TN)

Table 1 shows the number of rows test data predicted with the true and false classification. Thus, the accuracy of classification can be obtained from the sum of true positive and true negative divided by the total, with the following formula:

$$Accuracy\ (A) = \frac{TP+TN}{(TP+TN+FP+FN)} \tag{3}$$

4.6. Integration

This stage is the integration of the results of the implementation of the previous stages which is the developing of software integration tools from the stages of the data collection, preprocessing, feature selection and extraction, classification, validation and evaluation.

5. Conclusion

This research resulted the proposed framework for implementation of Sentiment Analysis of Indonesian Language Tweet on Twitter for Work Programs of the Bandung City Government. To implement this proposed framework, the development of software applications for each steps, from Data Collection stages, Preprocessing, Selection and Feature Extraction, Classification, Validation and Evaluation, and Integration is required.

Finally, This proposed framework is expected to be developed in the future research to make the application of implementation of Sentiment Analysis of Indonesian Language Tweet on Twitter for Work Programs of the Bandung City Government.

References

[1] Pang B, Lee L and Vaithyanathan S 2002 Thumbs up? Sentiment Classification using Machine Learning *in Proc. of the ACL-02 conf. on Empirical methods in natural language processing* vol 10 pp 79-86 (USA: Morristown, NJ)

[2] Pang B and Lee L 2008 Opinion Mining and Sentiment Analysis, Foundations and Trends *Information Retrieval* vol 2 no. Issue 1-2 pp 1-135

[3] Liu B 2010 Sentiment Analysis: A Multi-Faceted Problem *IEEE Intelligent Systems*

[4] Liu B 2010 Sentiment Analysis and Subjectivity *Handbook of Natural Language Processing*

[5] Boiy E, Hens P, Deschacht K, Moens M F 2007 Automatic sentiment analysis in on-line text *Proc. of the 11th Intl. Conf. on Electronic Publishing (ELPUB 2007*) (Austria: Vienna)

[6] Nur M Y and Santika D 2011 Analisis Sentimen Pada Dokumen Berbahasa Indonesia Dengan Pendekatan Support Vector Machine (Indonesia : KNSI 2011)

[7] O'Keefe T and Koprinska I 2009 Feature Selection and Weighting Methods in Sentiment Analysis *Proc. of the 14th Australasian Document Computing Symp.*

[8] Pak A and Paroubek P 2010 Twitter as a Corpus for Sentiment Analysis and Opinion Mining *Proc. of the Seventh conf. on Intl. Language Resources and Evaluation (LREC'10)* pp 1320-1326

[9] Naradhipa A R and Purwarianti Ayu 2011 Sentiment Classification for Indonesian Message in Social Media *Intl. Conf. on Electrical Engineering and Informatics* (Indonesia : Bandung)

[10] Vapnik V 2000 The Nature of Statistical Learning Theory *Springer-Verlag* pp 863-884

[11] Yang Y and Liu X 1999 A re-examination of text categorization methods *Proc. of ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR)* (USA : New York) pp 42–49

[12] Olson D L and Delen D 2008 Advanced Data Mining Techniques, 1st ed. Heidelberg (Berlin: Springer)