TELKOMNIKA, Vol.15, No.4, December 2017, pp. 1858~1864
ISSN: 1693-6930, accredited **A** by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v15i4.4633

Sentiment Mining of Community Development Program Evaluation Based on Social Media

Siti Yuliyanti*[1], Taufik Djatna[2], Heru Sukoco[3]
[1,3]Department of Computer Science, Bogor Agricultural University, Indonesia
[2]Department of Agro-industrial Technology, Bogor Agricultural University, Kampus IPB Darmaga P.O. Box 220 Bogor, (62-251)8621974/(62-251)8621974, Phone: (0251) 86228448, Fax: (0251) 8622986, Indonesia
*Corresponding author, e-mail: sitiyuliyanti@apps.ipb.ac.id[1], taufikdjatna@ipb.ac.id[2], hrskom@ipb.ac.id[3]

Abstract

It is crucial to support community-oriented services for youth awareness in the social media with knowledge extraction, which would be useful for both government agencies and community group of interest for program evaluation. This work provided to formulate effective evaluation on community development program and addressing them to a correct action. By using classification based SVM, evaluation of the achievement level conducted in both quantitative and qualitative analysis, particularly to conclude which activities has high success rate. By using social media based activities, this study searched the sentiment analysis from every activities comments based on their tweet. First, we kicked off preprocessing stage, reducing feature space by using principle of component analysis and estimate parameters for classification purposes. Second, we modeled activity classification by using support vector machine. At last, set term score by calculating term frequency, which combined with term sentiment scores based on lexicon.The result shows that models provided sentiment summarization that point out the success level of positive sentiment.

Keywords*: activities, evaluation, parameter, responses, sentiment*

1. Introduction

Opinion mining or sentiment analysis that is part of text mining process to extract and process textual unstructured data automatically to obtain sentiment information in a sentence. Opinions are analyzed to see trends, issues or object on their negative or positive perspective [1]. In this work, an extensive series of knowledge discovery processes were addressed problem soluition based on communication or action changes in the field of community development. Community development is a constructed movement that is designed to improve the overall living standard of the people through active participation and initiative of the community [2] [3].

Problems often arise in implementation and evaluation of community development programs such as relationship, structure, power, shared for meaning, communication for change, motivation to decision making and integration of disparate concerns [6]. Twitter as a means of delivering information in socialization activities, related to the issue of communication for change. Abundance of tweets can be analyzed to provide information about public sentiment, and measuring the level of activity for better dissemination of information.

We emphasize the important role of social media as a group of Internet-based applications that was built based on membership participation during sharing ideas and opinion. Particularly with advance of Web 2.0 technologies, which has enabled creation and exchange of user-generated content [4]. In addition to Brogan in 2010 on his book entitled “Social Media 101 Tactic and Tips to Develop Your Business Online” appointed that social media is a new set of communication and collaboration tools that enable many types of interaction that was not previously available in common. As an important example, Twitter is a microblogging site to enable users to send their tweet as maximum as 140 characters. Nowadays, there is many elements on Twitter content such as Profile, Following, Follower, Mentions (@), Direct Message, Hashtag, Trending Topics, and others. [5].

In terms of community development evaluation within Twitter transaction, how the opinionated postings on social media (e.g., reviews, forum discussions, blogs, micro-blogs,

Received July 20, 2017; Revised October 28, 2017; Accepted November 13, 2017

Twitters, comments, and postings on social network sites) have helped reshaping community member perspective, and sway public sentiments and emotions, which has affected profoundly on overall social impact of the program development. It is how sentiment analysis process as a computational study of people's opinions, sentiments, emotions, and attitudes works well within idea transformation during community development processes. This fascinating problem is increasingly important in building society capacity. It offers numerous research challenges but promises insight useful to anyone interested in opinion analysis and social media analysis [7].

For the purpose of evaluation, a collection of tweets was set. These tweets were extracted and processed for information such as sentiment in dedicated tweets of community development issues. Sentiment words can be divided into two types, base type and comparative type. All the preceding example words are of the base type. Sentiment words of the comparative type (which include the superlative type) are used to express comparative and superlative opinions. Examples of such words include better, worse, best, worst, and alikes, which are comparative and superlative forms of the base adjectives or adverbs such as good and bad. Sentiment analysis of tweets is used to find out whether a tweet consists of positive, negative or neutral sentiment.There are two kinds of learning that usually used in the process of sentiment analysis, which is supervised learning and unsupervised learning.

Several researchers have studied how to adapt a general sentiment lexicon to a particular domain. The machine learning method belongs to supervise learning; this method usually needs many training data that have been labeled manually. Without labeling the training data, supervise learning will disable to be processed [7]. The lexicon-based method is belong to unsupervise learning, which does not need training data and only depend on the dictionary that is used. Both methods have different characteristics, but it can complement if both methods are combined. Regarding to the techniques used in a sentiment analysis, there are two major techniques commonly used; those are machine learning based and lexicon based [8]. Supervised machines learning techniques that are commonly used for this purpose including support vector machine (SVM) [9] and Naïve Bayes [10].

The combination of both methods can be done by using lexicon-based method to create label tweets which can be used as training data in SVM method so there will be no manually labeling process in this combination method [11] [12]. SVM proved to provide a good classification result in sentiment mining, the implemented practically SVM is often far from the expected level theoretically because their implementations are based on the approximated algorithms due to the high complexity of time and space. Improving the limited performanceclassifications of the real SVM. PCA is deployed to decrease the complexity of an SVM-based sentiment classification task by applying the concept of reducing the data dimensionality.

Contrary to several previous research methods, which already investigatedopinion corpus was written mostly in the English, this work considered a community development opinion problem written in Bahasa Indonesia. This challenge obviously has different structure than did in English. By building a model to determine the sentiment of a tweet about the public responses to the activities of the community development program. Those tweets will be preprocessing, reduction feature and classify to find out whether tweets consists of positive, negative or neutral sentiment. This research focuses to activities of community development programs in the area of Bogor. First, crawling dataset about activities of community development and the preprocessing that is used in this research: filter, lower case, removal the stopwords, tokenizing, parsing, labeling sentiment, and weighting terms. Second, the feature with the lowest value of principal component is reduced using PCA to facilitate the classification using SVM. Finally, evaluation models using scenario testing on comparisons models of training data and test data that are used.

2. Research Method

Dataset of this case study was written using Bahasa related to the activities of the community development program that previously observed directly in the office of Bogor municipial governmental agency, West Java Indonesia. We collected more than 2000 tweets from twitter about two prominent youth awareness activities. As shown in Figure 1, the research framework was divided into 4 parts: those are data collected, pre-processing, classification sentiment and evaluation. We discussed the details and result in the following section.

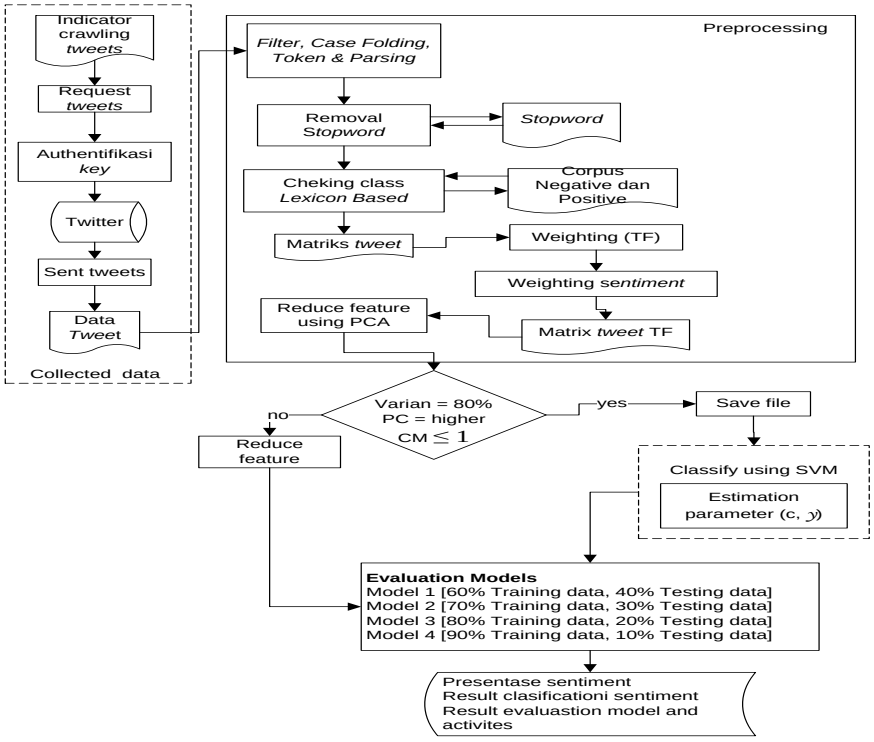

Figure 1. Research Framework

2.1. Preprocessing

For the purpose of pre-processing on the data set, the following sub processed initiated by filter, lower case, tokenize, remove the stop-words, weighting and class labeling were applied. As the pre-processing stages completed and resulted to data that had been cleaned, and then the later one will be processed in by weighting words, with calculation of term frequency-inverse document frequency (TF-IDF) and labeling sentiment using lexicon-based method. Lexicon-based is belong to unsupervised techniques; this method classifies the data into 2 classes of positive or negative [8], [10]. This lexicon-based method adapts the word-level polarities of a general-purpose sentiment lexicon for a particular domain by utilizing the expression-level polarities in the domain. In return, the adapted word-level polarities are used to improve the expression-level polarities.

The word-level and the expression-level polarity relationships are modeled as a set of constraints and the problem is solved using integer linear programming. This method is based on the help of dictionary to classify the tweet into positive sentiment, negative sentiment, or neutral sentiment. There are several steps of lexicon-based that is used in this research, such as determining the polarity of words, negation handling, and also giving a score to every each entity in the tweet. The formula to calculate the score for the entity as seen in the formula (1), based on [7].

$$\text{score}(f)= \sum_{\omega_i:\omega_i \in S \cap \omega_i \in V} \frac{\omega_i . s_0}{\text{distance}(\omega_i, f)} \quad (1)$$

where:
score(*f*)= The final label score of the feature

ω^i = An sentiment word
S = All sentiment words
V =sample space feature and sentiment word
s_0 = Label of the sentiment word (+1,0, or -1)

distance ($_{\omega^i}$,*f*)= Distance between feature (*f*) and the sentiment words (w_i)

In general, calculation of TF (Term Frequency) is the calculation of the number of times a word against the tweet. It is to show how important a word to a tweet that there is a collection of tweets [17]. Results of phases D and E used as a vector W. where W={w1, w2, …, wi} and i ε s contain the word candidates sentiment and W ε V with V is a corpus that contains features and word sentiment. This step gives the class label with lexicon based on each tweet by positive and negative classes that exist in the Indonesian lexicon corpus. Furthermore, the proximity value is calculated by using the lexicon corpus Equation 1. If the value is positive or end score then assumed the features is apositive. Then tweet or end score value is negative, it is assumed negative features in the tweet grudges, and if not both then tweet including neutral class [7]. TF-IDF [11] was used to identify how important is every single available term in the corpus. It is also a common technique to calculate the vector weight based on the semantic relatedness, *tft,d*, the frequency of term *t* in document *d*, is defined as formula (2). In this work document is a tweet and used just term frequency.

$$tf_{t,d} = \frac{f_{t,d}}{\arg\max(tf_d)} \quad (2)$$

2.2. Feature Reduction

The purpose of dimension reduction is the main task of PCA (Principal Component Analysis) algorithm [15]. In this work, PCA is used as a statistical procedure in exploratory data analysis and for making predictive opinion models. PCA can be computed by eigenvalue decomposition of a data covariance (or correlation) matrix or singular value decomposition of a data matrix, usually after mean centering (and normalizing or using Z-scores) the data matrix for each attribute.The results of a PCA are usually discussed in terms of component scores, sometimes called factor scores (the transformed variable values corresponding to a particular data point), and loadings (the weight by which each standardized original variable should be multiplied to get the component score). These principal components are used as a predictor or criterion variable in other analysis.

The variables are orthogonalized by the PCA and principal components with largest variation are chosen and components with least variation are eliminated from the dataset. PCA is powerful with its simplicity of the true eigenvector-based multivariate analyses in which the operation can be thought of as revealing the internal structure of the data in a way that best explains the variance in the data. If a multivariate dataset is visualised as a set of coordinates in a high-dimensional data space (1 axis per variable), PCA can supply the user with a lower-dimensional picture, a projection of this object when viewed from its most informative viewpoint. Users might interpret the result by using only the first few principal components so that the dimensionality of the transformed data is reduced.

2.3. Classification Methods

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.

SVM are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. 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 [17]. 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, Term Presence, etc.) will result in a value or pattern that will be used in the testing process for labeling process tweet.

SVM help to find the optimal hyperplane that has a maximum margin and constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. Margin itself is the distance between the hyperplane (lines) with the closest point from each class, the closest point usually called as support vector. This section discusses the classification methods used in this research to develop the sentiment mining models. Classification was compiled byRapidMiner Studio with default values for all parameter. The steps SVM of conducted in this research are as follows [17] to prepare the input data in the form of index word, the weight of word, and also its label; to calculate the parameter weight *(w); to*. Calculate the bias (*b*) and to get the classification for data testing the processes of four steps performed on SVM method in this research using by setting the best parameter estimation using the grid search. Grid search aims to make the grid parameters of each pair (C, γ). Parameter values (C, γ) determined in advance by a range of values from 0.1 to 0.9, and then pair each value of the parameters (C, γ) so the couple parameters that yield the highest accuracy used in testing scenarios 4 models based on the percentage of training data and test data. The scenario tests the comparisons of training data and testing data that are used [7]. The detail of the comparison of both these data are as follows:

a. First model with overall training data 60% and the rest as testing up to 40%
b. Second model with overall training training data 70% and testing data 30%
c. Third nodel with overall training training data 80% and testing data 20%
d. Fourth model with overall training training data 90% and testing data 10%

The testing scenario was gained from the data tweet that has successfully analyzed. This testing model showed the conclusion on the amount of positives, negatives and neutral sentiments obtained from each activities of community development program. The conclusion was only taken from the testing data that previously got the highest accuracy in each activities of community development program.

3. Results and Analysis

3.1. Preprocessing

Crawling is the process by means of a registration API connection using Twitter Application Management to get the API Key, API Secret, access token, access token secret and then performs authentication. Further data collection by keyword with the desired parameters, for example, in this study keywords used about the activities within one years of January 2015 until January 2016 and then stored in a file with the .csv (comma delimited). After preprocessing of the dataset that includes: filter, lower case, tokenize, remove the stop-words, weighting and class labeling.

3.2. Feature Reduction

The reduction process used to find the features of the best features that will be used in a classification process that is using the PC of the highest value and reducing features that are considered unfavorable. Features shown are featured with the highest value PC with cumulative variance≤1 that means that features that do not meet these values are no longer variants. Based on reduction features, 1219 feature of activities 1 and features 951 feature of activities 2 that will be used in the classification process.

3.3. Classification Performance

In this study, prior to classification by test data, will be estimation parameters on SVM to find the best parameter to be used for classification are the parameter c and γ. SVM classification process in the present study using RBF kernel function where the kernel requires parameter c and γ at the process [18]. During the process to get the best parameter values, it is conducted several stages on the dataset. The first has done by creating a grid parameter on each pair of parameter values. Parameter values c and γ predetermined manually with a range

of values each 0.1 to 0.9. Couple grades c and γ the best is that the average value of the most accuracy, some couples parameter values that provide accuracy best in class classification sentiment amounted to 97.44% is (c=0.8, γ =0.8), (c=0.8, γ=0.9), (c=0.9, γ=0.8) and (c=0.9, γ=0.9). Application of SVM algorithm with the addition of a neutral class is expected to produce a good model with a high degree of accuracy. Illustrations classification process is represented in Figure 2.

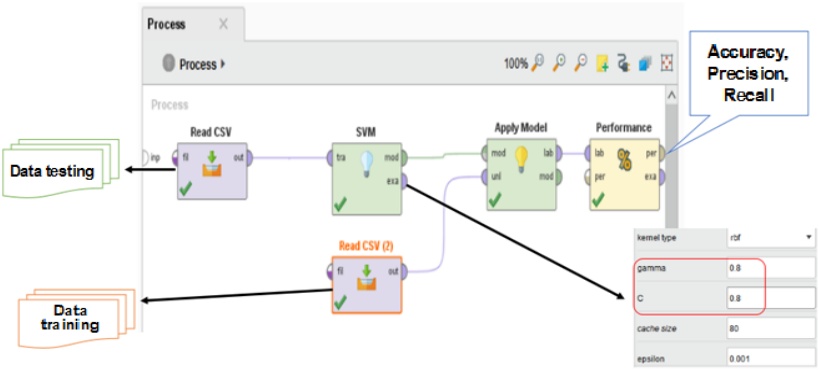

Figure 2. Flow knowledge classification sentiment

4. Evaluation

Doing performance the classification task, the class evaluation was done by using test scenario use result estimation parameter. It would be easy to understand how good a model in classification in Table 1. That divides the data into four models: model 1 with 60% of training data, 40% of test data, model 2 with 70% of training data, 30% of test data, model 3 with 80% of training data, 20% of test data, and model 4 90% 10% training data test data to determine the level of accuracy of the model. Table 1 shows that the highest accuracy in the classification by using the dataset obtained by reduction feature of the Model 3 for the activities of the Activity 1while for activities Activity 1in Model 1. Evaluate the performance of the classification model is based on three parameters: accuracy, precision, and recall the values indicated means in Table 1, where accuracy is not significantly affected by the division of training data and test data. This level of accuracy is good enough to compare with previous research [13] and by comparing the results of classification without reduction feature by PCA.

Reduction feature can improve the accuracy of the classification process and know the public response to the activities of the community development program through Twitter. The weakness of this study is not done preprocessing to detect the language of 'Alay' in a tweet, has not presented a sentiment classification spatially group activities and social media are used in the extraction only the Twitter dataset.

Table 1. Accuracy of Classification Sentiment using Parameter c=0.8 and *γ=0.8*

Model	Acitivity 1	Acitivity 2
First Model (60% data training; 40% data testing)	82.78	88.64
Second Model (70% data training; 30% data testing)	79.49	85.35
Third Model (80% data training; 20% data testing)	78.75	85.71
Fourth Model (90% data training; 10% data testing)	78.39	83.15
Mean79.85		85.71

6. Conclusion

Sentiment mining models are built which capable for extraction textual data into structure so that produce sentiment and classified to determine the public response to the activities in community development programs. Data collects on the crawling tweets in 1000 tweets of each activity, after preprocessing feature obtained in 1219 and 1302 features and reduced feature after feature into 1156 and features 951. The couples parameter values that

provide best accuracy in class classification sentiment is amounted to 97.44% is (c=0.8, γ=0.8), (c=0.8, γ=0.9), (c=0.9, γ=0.8) and (c=0.9, γ=0.9). The accuracy of the resulting data is tweet by reduction features highest Model 1 with 60% of training data and 40% of test data on the Activity 2, namely an accuracy of 88.64% and 82.78% of activity 1. The level of accuracy of the model affected SVM parameter estimation and preprocessing, but not affected the distribution of test data and training data. The evaluation program is a Activity 2 have a level of information spread with the best positive sentiment, and then Activity 1 should be increased of spread information and dissemination program of activities.

References

[1] Pang B, Lee L. Opinion Mining and Sentiments Analysis. *Foundations and TrendsR in Information Retrieval*. 2008; 2(1-2): 1-135.
[2] Rahman R. Corporate Social Responsibility. Yogyakarta: Media Pressindo. 2009.
[3] Adi IR. Empowerment, Community Development, *dan Intervensi Komunitas*: *Seri Pemberdayaan Masyarakat* 03. Publisher Institution Faculty of Economics University Indonesia. 2003. ISBN: 979-9242-44-5.
[4] Hemalatha I, Varma PG, Govardhan A. Preprocessing the Informal Text for Efficient Sentiment Analysis. *IJETTCS*. 2012; 1. ISSN: 2278-6856.
[5] Ho C, Pong L. Interpreting TF-IDF Term Weights as Making Relevance Decision. ACM. 2008.
[6] Philips R, Pittman RH. *An Introduction to Community Development*. ISBN: 0-203-88693-3. First published by Routledge, USA and Canada. 2009.
[7] Tiara, Sabariah MK, Effendy M. Sentiment Analysis on Twitter Using the Combination of Lexicon-Based and Support Vector Machine for Assessing the Performance of a Television Program. ICoICT. 2015.
[8] Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*. 2014: 1093-1113.
[9] Xu K, Shaoyi S, Li J, Yuxia S. Mining comparative opinions from customer reviews for Competitive Intelligence. *Decision Support Systems*. 2011; 50: 743-754.
[10] Li N, Wu DD. Using Text Mining and Sentiment Analysis for Online Forums Hotspot Detection. *Decision Support Systems*. 2010; 48: 354-368.
[11] Tan S, Wang Y, Cheng X. *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.2008. Singapore.
[12] Pang B, Lee L, Vithyanathan S. *Thumbs Up? Sentiment Classification using Machine Learning Techniques*. Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing (pp. 79-86). 2002. Stroudsburg: Association for Computational Linguistic.
[13] Subramanian KM, Venkatachalam K. Framework for Evaluating Camera Opinions. *Research Journal of Applied Sciences, Engineering and Technology*. 2015; (7): 519-525. ISSN: 2040-7459; e-ISSN: 2040-7467.
[14] Wahyudin I, Djatna T. Cluster Analysis for SME Risk Analysis Documents Based on Pillar K-Means. *TELKOMNIKA*. 2016; 14(2): 674~683. ISSN: 1693-6930.
[15] Jotheeswaran J, Loganathan R, Madhu SB. Feature Reduction using Principal Component Analysis for Opinion Mining. IJCST. 2012; 3(5): 118-121. ISSN 2047-3338.
[16] Vinodhini G, Chandrasekaran RM. Opinion Mining using Principle Component Analysis Based Ensemble Model for E-Commerce Application. CSIT. 2014; 2(3):169-179. DOI 10.1007/s40012-014-0055-3. Spinger.
[17] Putranti ND, Winarko E. *Analisis* Sentiment Twitter *untuk Teks Berbahasa Indonesia dengan* Maximum Entropy *dan* Support Vector Machine. *IJCCS*. 2014; 8(1): 91-100. ISSN: 1978-1520.
[18] Muis IA, Affandes M. *Penerapan Metode* SVM menggunakan Kernel Radial Basis Function (RBF) pada Klasifikasi Tweet. *Journal of Science Technology and Industry.* 2015; 12(2): 189-197. ISSN: 1693-2390.