Sentiment Analysis Using Multinomial Logistic Regression

Ramadhan WP[1], Astri Novianty S.T.,M.T[2],Casi Setianingsih S.T.,M.T[3]
[1,2,3]Department of Computer Engineering, Telkom University
Bandung, Indonesia

[1]ramaprakoso@students.telkomuniversity.ac.id, [2]astrinov@telkomuniversity.ac.id ,
[3]setiacasie@telkomuniversity.ac.id

Abstract- **Data amount becomes rapidly increased in today's era. Data can be in form of text, picture, voice, and video. Social media is one factor of the data increase as everybody expresses, gives opinion, and even complains in social media.**

The first step is data collection used API twitter with each candidate names on Jakarta Governor Election. The collected data then became input for preprocessing step. The next step is extracted-each tweet's feature to be listed. The list of features were transformed into feature vector in binary form and transformed again used Tf-idf method. Dataset consists of two kinds of data, training and testing. Training was labeled manually. K-Fold Cross Validation is used to test algorithm performance.

Based on the result of the test, accuracy obtained reached 74% in average with composition of training data and testing data by 90:10. Changed folding amount gave no impact to the accuracy level.

Keyword : twitter, multinomial logistic regression, text mining, softmax regression

I. INTRODUCTION

Social media have a lot of influence in social life. During this time, people would rather express his feeling on social media. Sentiment analysis or so-called opinion mining is a field of study that analyzes opinions, sentiments, evaluations, judgments, attitudes, and emotions on an entity such as products, services, organizations, individuals, problems, events, topics and their attributes [1]. Many methods are used to perform sentiment analysis such as Naive Bayes, Maximum Entropy, Support Vector Machine and others [7]. Machine learning techniques use a set of training and test set for classification [7]. According to Pang [3], the method of using machine learning can provide better results, but the classification of supervised learning require a lot of training data that has been labeled. Without labeled training data, supervised learning can't work [4]. The selection of the Multinomial Logistic Regression method is chosen because competitive in terms of CPU and memory consumption [9]. Multinomial Logistic Regression is preferred when we have features of different type (continuous, discrete, dummy variables etc.), nevertheless given that it is a regression model, it is more vulnerable to multicollinearity problems and thus it should be avoided when our features are highly correlated [9].

II. LITERTURE REVIEW

A. Twitter

Twitter is an online social media and microblogging service that lets users post, send and read text-based messages of up to 140 characters. Twitter was founded by Jack Dorsey in 2006 and became one of the most popular social media services in the world. Twitter experienced a fairly rapid growth until January 2013 the number of users reached the number 500 million registered on twitter, 200 million of them are active users. In Indonesia alone according to Dick Costolo during a visit to Indonesia in 2014 Twitter users in Indonesia reached 50 million, based on the latest data 2014 active users Twitter 284 million. In June 2012 Jakarta ranks first with 10.6 million tweets [2].

B. Sentiment Analysis

The Sentiment analysis, usually called as opinion mining, is a determination of the emotions behind a series of words, which are used to gain an understanding of attitudes, opinions and emotions expressed online. The sentiment analysis is very useful in monitoring social media because it allows us to gain a picture of the broader public opinion of a particular topic.

C. Feature Extraction

Feature extraction is a process of transforming data input into a feature set [6]. A feature is an object of a pattern whose quantity can be measured, its classification by virtue of each of these features. The ability of the machine learning process is highly dependent on its features so it is important to choose the purpose of feature extraction [5].

978-1-5386-1667-3/17/$31.00 ©2017 IEEE

D. Multinomial Logistic Regression

Multinomial Logistic regression, also known as Softmax Regression due to the hypothesis function that it uses, is a supervised learning algorithm which can be used in several problem including text classification [9]. It is a regression model which generalizes the logistic regression to classification problems where the output can take more than two possible values [9]. Training dataset consist of m (x_i,y_i) pairs and let k be the number of all possible classes. Also by using the bag-of-word let {w_1,...,w_n} be the set of n words that can appear within our texts[9].

E. Validation and Evaluation of Sentiment Classification

To validate the accuracy of the system used k-fold method and to measure the validity of the tweet is used confusion matrix. In k-fold cross validation, the initial data is randomly partitioned into a subset (fold), each of the same size. The training and testing process is done as many times as [7]. Table I shows the confusion matrix which is used to assist in calculation of the evaluation system[8].

Table I Confusion Matrix [8]

	Predicted Positive	***Predicted Negatives***
Actual Positive Instance	*Number of True Positives Instance* (TP)	*Number of False Negetives Instances* (FN)
Actual Negative Instance	*Number of False Positives Instance*(FP)	*Number of True Negatives Instance*(TN)

The formula that used in the confusion matrix configuration [8]:

1. Accuracy
   The value of the entire true predicted against all predicted. The formula to obtain accuracy can be seen in equation.
$$Accuracy = \frac{TN+TP}{TN+TP+FP+FN}$$
2. Precision
   The value of true positive prediction againts all positive prediction. The formula to obtain precision be seen in equation.
$$Precision = \frac{TP}{TP+FP}$$
3. Recall
   The value of true positive prediction against all actual positive. The formula recall can be seen in equation.
$$Recall = \frac{TP}{TP+FN}$$

III. METHODOLOGY

This stage will give explanation of the process. Before that try to look at the general process in this system in Figure 1.

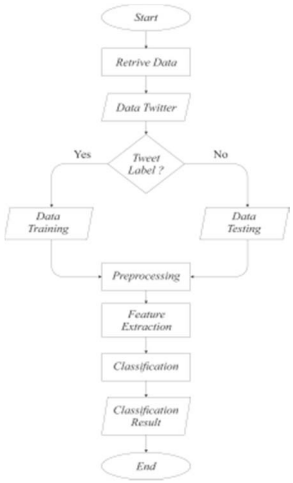

Fig 1 General process of the System

Details the process description :

1. Retrieved of raw data in the form of tweet using API from twitter then data stored in csv form. For labeled data is training data and unlabeled data is data testing.
2. In the data testing or training data is done preprocessing. The process of preprocessing involves deleting URLs, punctuation, deleting stop words, changing the word slang to raw and stemming.
3. Then performed feature extraction process on tweet that has been clean result of preprocessing. The feature extraction process includes word grouping with the Bag Of Words method and feature weighting with Tf-idf.
4. Tweet that is already a collection of features in the Bag of Words and has been given a weighting using Tf-idf classified using the Multinomial Logistic Regression method.
5. The classification process produces a recall value, precision, accuracy and sentiments of tweet.

IV. TESTING

A. Dataset

In doing the testing, the dataset is divided into 2 data training and data testing. For the category, the training and testing data contains mixed data between data from pairs a, b, and c, combined into one. Then the data is combined into one then divided according to the scenario. In the test data system is labeled manually while the test data is not labeled, so the system can predict whether the value of test positive or negative tweet. To calculate the performance of the system using a system of cross validation, to assess or validate the accuracy of a model built on a particular dataset.

Table II Details of Dataset

Data	Data Training Amount	Data Testing Amount
50%:50%	753 tweet	753 tweet
60%:40%	904 tweet	904 tweet
70%:30%	1054 tweet	1054 tweet
80%:20%	1205 tweet	1205 tweet
90%:10%	1356 tweet	1356 tweet

Table III Number Sentiment label of Training Data

Data Composition	Sentiment	
	Positive	Negative
50%:50%	361 tweet	392 tweet
60%:40%	332 tweet	462 tweet
70%:30%	522 tweet	532 tweet
80%:20%	598 tweet	607 tweet
90%:10%	682 tweet	672 tweet

B. Testing Scenario :

1. The first test uses a comparison of data composition between training and testing data..
2. The second test because cross validation uses k-fold, the test is divided into 10-Fold. For Multinomial Logistic Regression method the data used is the most optimal composition of the first. From the test is searched fold to how with the highest accuracy, precision and recall value.

C. Testing Result

Analysis In this section will be presented the results of testing and analysis of the results that have been obtained from the testing process.

- **Analysis based on comparison of training data and testing data**

The first scenario is tested with composition data comparison training and data testing. From the comparison of data composition will be seen the most optimal.

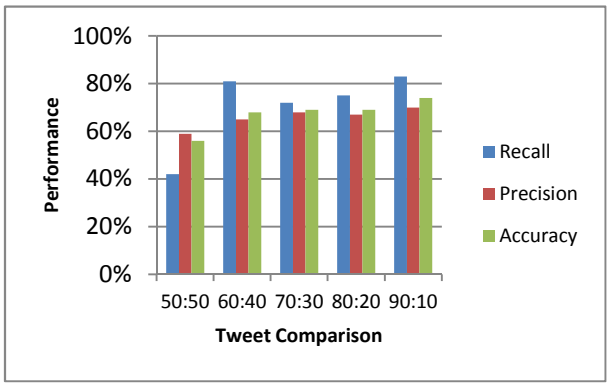

Figure II Result based on comparison of data

In the test, Multinomial Logistic Regression algorithm with a composition ratio of 90:10 has the most optimal result reaching 74% accuracy value. From the test can be seen the more the amount of training data, the higher the calculation results.

- **Analysis Based on Number of K-Fold**

In the second analysis, the value of fold is changed. The value of the fold is changed from 2-20 fold. For analysis on Multinomial Logistic Regression

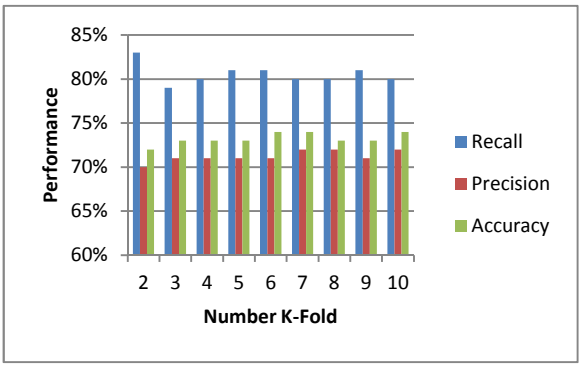

Figure IV Result based on number of K-Fold

From the test results, the greater the value of the fold affect the acquisition accuracy of each method. Multinomial Logistic Regression can achieve the highest accuracy up to 74%.

V. CONCLUSION

Based on the results of tests and analysis that have been done before, then drawn some conclusions as follows:

1. From all experiments, the most influential parameter in getting good result is the composition of training and testing data. The more training data compared to the amount of data testing, the higher the accuracy obtained.
2. The number of folding does not affect the performance of each method.

References

[1] Liu, B.*Sentiment Analysis and Opinion Mining*.Morgan & Claypool Publishers.

[2] "Semiocast — Geolocation Of Twitter Users (July 2012)". *Semiocast*. N.p., 2017. Web. 4 Mar. 2016.

[3] 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

[4] McCallum, A., Freitag, D., dan Pereira, F.2000. *Maximum Entropy Markov Models for Information Extraction and Segmentation*. Proc. ICML 2000, pp. 591–598, Stanford, California.

[5] Zainuddin, Nurulhuda and A. Selamat.2014. Sentimen Analysis Using Support Vector Machine".*International Conference on Computer, Communication, and Control Technology*.

[6] W. B. Croft, D. Metzler, and T. Strohman, *Searcssh Engines Information Retrieval in Practice*. Addison Wesley, 2009.

[7] 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.

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

[9] Vryniotis, Vasilis, and Vasilis Vryniotis. "Machine Learning Tutorial: The Multinomial Logistic Regression (Softmax Regression) | Datumbox". *Blog.datumbox.com*. N.p., 2017. Web. 24 May 2017.