IOP Conf. Series: Materials Science and Engineering **288** (2018) 012162 doi:10.1088/1757-899X/288/1/012162

Classification of Indonesian quote on Twitter using Naïve Bayes

A Rachmadany[1*], Y M Pranoto[2], Gunawan[2], M T Multazam[3], A B D Nandiyanto[4], A G Abdullah[5] and I Widiaty[6]

[1]Fakultas Teknik Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia
[2]Sekolah Tinggi Teknik Surabaya, Indonesia
[3]Fakultas Hukum, Universitas Muhammadiyah Sidoarjo, Indonesia
[4]Departemen Kimia, Universitas Pendidikan Indonesia, Jl. Dr. Setiabudi No. 229, Bandung 40154, Jawa Barat, Indonesia
[5]Departemen Pendidikan Teknik Elektro, Universitas Pendidikan Indonesia, Jl. Dr. Setiabudi No. 229, Bandung 40154, Jawa Barat, Indonesia
[6]Departemen Pendidikan Kesejahteraan Keluarga, Universitas Pendidikan Indonesia, Jl. Dr. Setiabudi No. 229, Bandung 40154, Jawa Barat, Indonesia

*rachmadany@umsida.ac.id

Abstract. Quote is sentences made in the hope that someone can become strong personalities, individuals who always improve themselves to move forward and achieve success. Social media is a place for people to express his heart to the world that sometimes the expression of the heart is quotes. Here, the purpose of this study was to classify Indonesian quote on Twitter using Naïve Bayes. This experiment uses text classification from Twitter data written by Twitter users which are quote then classification again grouped into 6 categories (Love, Life, Motivation, Education, Religion, Others). The language used is Indonesian. The method used is Naive Bayes. The results of this experiment are a web application collection of Indonesian quote that have been classified. This classification gives the user ease in finding quote based on class or keyword. For example, when a user wants to find a 'motivation' quote, this classification would be very useful.

1. Introduction

The Internet has become one part of everyday life, so information can be quickly obtained through the internet. Various media have been built using internet communication network. One of the media sources of information is blog or web log that is managed by individuals or groups of authors containing various informations. The type of micro blog is a variation of the blog which contains less data than the blog [1].

Twitter which is one of the famous micro blogs in a social networking service [2-4]. Twitter allows users to share information in the form of short text 140 characters [5]. Twitter users are very large and from various circles, each user can subscribe to receive messages from other users by becoming 'followers'. Twitter has played an important role in providing information to users and also has changed the way people convey information one of which is the delivery of quotations or wisdom sentence. Quote is made in the hope that someone can become strong individuals, individuals who always improve themselves to move forward and achieve success. The quote is written and conveyed in order to benefit

the readers, so they will always be the ones who are always advancing toward success without losing themselves [6].

However, a quote sentence posted on Twitter requires an effort to find it. Although searching with keywords on Twitter, the results of tweets in the show is not all a quote sentence [4]. Users should read the tweets one by one to determine whether the tweet is a quote sentence or not. Users also need to select whether the tweet is a quote phrase in accordance with the sought for example want to find the quote sentence about love, life, motivation, religion, and other. Thu, the purpose of this study was to classify Indonesian quote on twitter using Naïve Bayes. To alleviate this problem and to convey quote data effectively, this research is expected to produce web applications [7]. The web application that can collect and classify Indonesian quote sentences from data obtained through Twitter using the Naïve Bayes classification method.

2. Experimental Method

The input data used is Twitter tweet, taken using the Twitter Stream API [4]. Twitter API getting tweet contain quote or quotes and language Indonesian. From the data will be taken at random 1700 data tweet, then taken 1500 data used as training data and 200 data as testing data. For further training data, manual classification by human beings based on its benefits quote or not. If the tweet is quote sentence then classification again for each tweet to category in the type of quote sentence (Love, Life, Motivation, Religion, Education, Others). The data will be used as the basis to be processed by the system using naïve bayes algorithm as training data. The formula for calculating the class of the new tweet is shown by the following equation:

$$P_{(c)} = \frac{N_c}{N} \tag{1}$$

$$P_{(w|c)} = \frac{count(w,c)+1}{count(c)+|v|} \tag{2}$$

$$P_{(c|d)} = log(P_{(c)}) \times \prod_{k=n}^{p} log(P_{(w|c)}{}^{k}) \tag{3}$$

where P, c, N, w, v, d, and k are the probabilities, the class, the count, the word, the unique word, the word n, and the word count.

Before using Naïve Bayes algorithm, training data tweet must be pre-processing in covering stages. Tokenizer as the tokenizer process in Twitter has a difference with the tokenizer process in other text. This is because the emoticons are often used by users. Tokenizer stage starts from separating the tweet section separated by a space character. Furthermore, sections that have only one non-alphabetical character, numbers, symbols and emoticons will be discarded. Normalization, there are some typical components that usually exist in tweet, username, URL, “RT” (retweet), and hash tags. Username, URL, and “RT” do not have any effect, then the three components will be discarded. The username component is identified by the appearance of the “@” character, hash tags identified by “#” while the URL component is recognized via regular expression (regex). Case Folding, for more effective all letters will be made lowercase. Clean Number, some tweets are containing number. The addition of frequent figures becomes a very significant influence, so the numbers on the tweet will be deleted. Stopword Removal, this process will eliminate the words that often appear but does not have any effect in the extraction of a tweet classification. Words that include this word are "the", "and", "in", "from" and so on. Stemming, get the basic word from the filtering result. At this stage the process of returning various forms of words into a similar representation [8,9].

After the pre-processing next step modelling the analysis using Naïve Bayes machine learning method for model formation. The Naïve Bayes classification method was chosen because of its easy use, simple design and complex problem-solving capabilities, Naïve bayes is a machine learning method that uses probability models or opportunities [6]. Training data in the form of tweets and class pairs

serve as the source of analysis model formation. Each feature that represents a tweet is calculated the probability of occurrence in a class tweet of a quote sentence or tweet that not a quote sentence, if the sentence is quote sentence then represented again the probability of each document for each category of quote sentence (Love, Life, Motivation, Religion, Education, Others) [7].

After the training data is processed with the naïve bayes algorithm, the system is ready and will automatically capture tweets from Twitter every minute using the Twitter API. Figure 1 shows the web server automatically every minute taking tweet data from the Twitter server using Stream API, each new data will be done pre-processing then calculated the probability value of each feature for each class. The greatest probability value of the calculation result is a new tweet class.

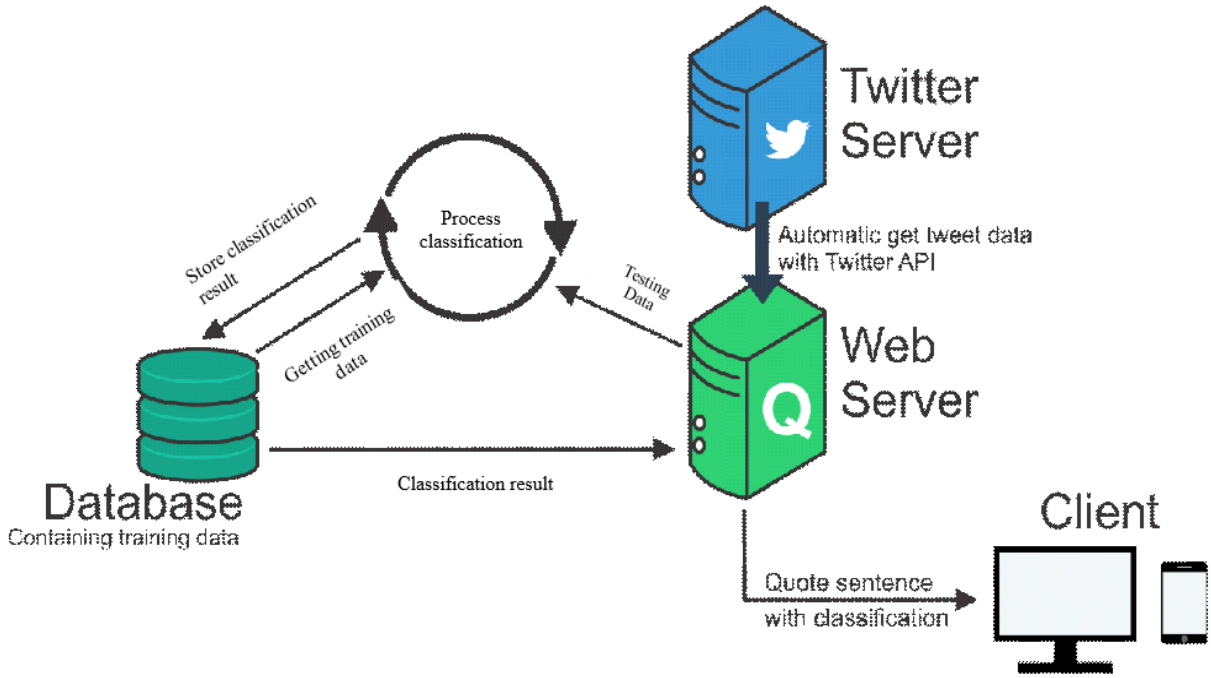

Figure 1. Work flow

3. Results and Discussion

Figure 2 shows the preprocessing of making the training data which the process includes (Tokenizer, Normalization, Case Folding, Clean Number, Stopword Removal, Stemming). Preprocessing work on every document or every tweet.

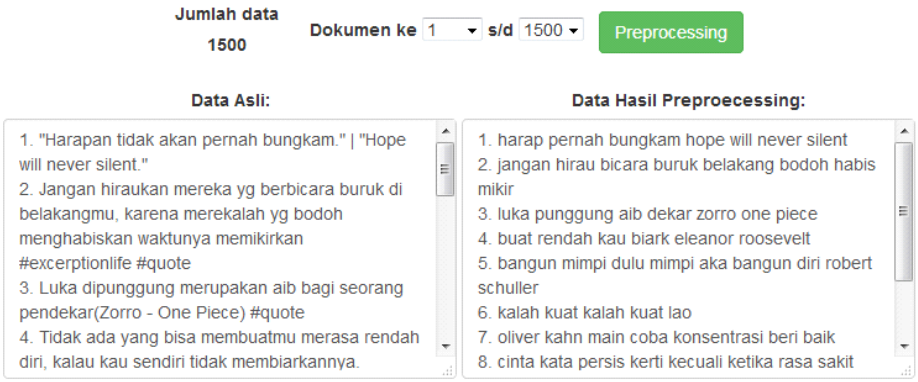

Figure 2. Preprocessing

Figure 3 shows the result process of Naïve Bayes algorithm with probabilities result, every word is calculate with other word in every document. In this process, every result probabilities store to array with json encode.

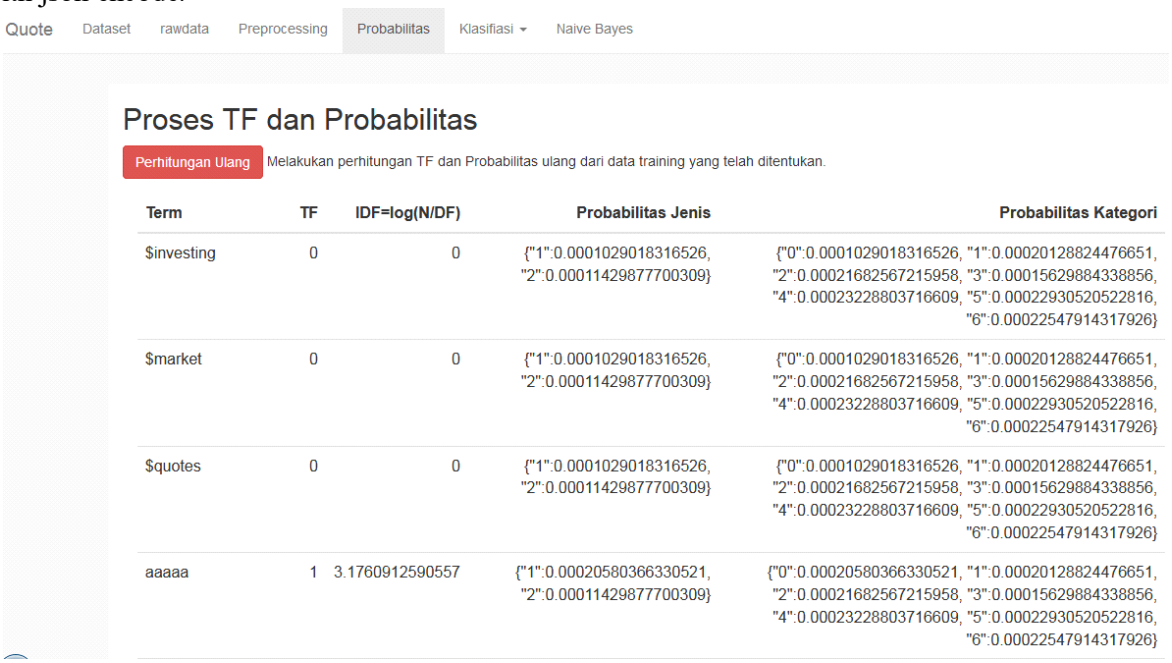

Figure 3. Creating Naïve Bayes Probabilities

Figure 4 shows the result proccess classification of Naïve Bayes algorithm to public user,tweets that are displayed to the user are tweets that have been classified and tweets that category quote sentence(Love, Life, Motivation, Religion, Education, Others).

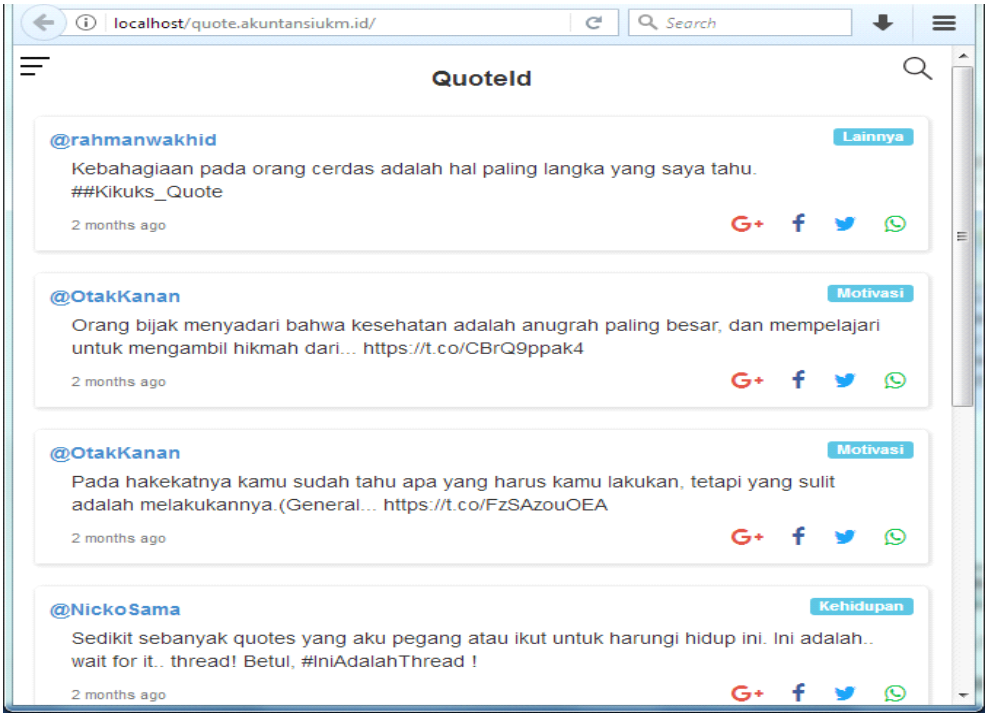

Figure 4. Classification result web application

The above results confirm that the present study was effective to classify quote number from Indonesia. The results of this experiment also gave a web application collection of Indonesian quote that have been classified. This classification gives the user ease in finding quote based on class or keyword. For example, when a user wants to find a 'motivation' quote, this classification would be very useful [10,11].

4. Conclusion

This study classifies tweet data taken from twitter whether the tweet includes a quote sentence or not. Tweets will be grouped into classes (Love, Life, Motivation, Religion, Education, Others). Using a naive bayes classification gets an average accuracy of 80%. This classification gives the user ease in finding quote based on class or keyword. For example, when a user wants to find a 'motivation' quote, this classification would be very useful.

Acknowledgements

We acknowledged to University Muhammadiyah Sidoarjo and Sekolah Tinggi Teknik Surabaya.

References

[1] C Firdaus, W Wahyudin and E P Nugroho 2017 Monitoring System with Two Central Facilities Protocol *Indonesian Journal of Science and Technology* ***2*** 1 8-25

[2] H Kwak, C Lee, H Park and S Moon 2010 What is Twitter, a social network or a news media? *Proceedings of the 19th international conference on World wide web* 591-600 ACM

[3] A Java, X Song, T Finin and B Tseng 2007 Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis* 56-65 ACM

[4] M Cha, H Haddadi, F Benevenuto and P K Gummadi 2010 Measuring user influence in twitter: The million follower fallacy *Icwsm* **10** 10-17 30

[5] R Dayani, N Chhabra, T Kadian and R Kaushal 2016 Rumor detection in twitter: An analysis in retrospect *Int. Symp. Adv. Networks Telecommun. Syst. ANTS* 4–6

[6] A Søgaard 2012 *Mining wisdom 1st Work. Comput. Linguist. Lit. (CLfL 2012)* 54–58

[7] C Tseng, N Patel, H Paranjape, T Y Lin and S Teoh 2012 Classifying twitter data with Naïve Bayes Classifier *2012 IEEE Int. Conf. Granul. Comput.* 294–299

[8] A Go, R Bhayani and L Huang 2009 Twitter Sentiment Classification using Distant Supervision *Processing* **150** 12 1–6

[9] B Y Pratama and R Sarno 2016 Personality classification based on Twitter text using Naive Bayes, KNN and SVM *Proc. 2015 Int. Conf. Data Softw. Eng. ICODSE 2015* 170–174

[10] S Prasad 2010 *Micro-blogging Sentiment Analysis Using Bayesian Classification Methods*

[11] E C Goncalves 2014 NBBR: A baseline method for the evaluation of bayesian multi-label classification algorithms *Proc. - 14th Int. Conf. Comput. Sci. Its Appl. ICCSA 2014* 245–247