Syaifudin & Puspitasari, 2017

Volume 3 Issue 1, pp. 110 - 122

Date of Publication: 30th January, 2017

DOI- https://dx.doi.org/10.20319/mijst.2017.31.110122

This paper can be cited as: Syaifudin, Y. W., & Puspitasari, D. (2017). Twitter Data Mining for Sentiment Analysis on Peoples Feedback against Government Public Policy. MATTER: International Journal of Science and Technology, 3(1), 110 - 122.

TWITTER DATA MINING FOR SENTIMENT ANALYSIS ON PEOPLES FEEDBACK AGAINST GOVERNMENT PUBLIC POLICY

Yan Watequlis Syaifudin
Information Technology Department, State Polytechnic of Malang, Indonesia
yan_ws@yahoo.com

Dwi Puspitasari
Information Technology Department, State Polytechnic of Malang, Indonesia
dwi_sti@yahoo.com

Abstract

Government policies often get positive or negative response from the public. The response from the community feedback can be conveyed through print and electronic media. With the rise of social media today, people have a tendency to convey such feedback through social media such as Facebook, Twitter, Instagram, Path and other social media. Thus, to determine the public response to this policy that has been implemented, the government needs to know how your feedback from people who come from social media. But because of the feedback, it is difficult to detect how many positive or negative response from the public. Therefore, in this study will develop a system to obtain data in the form of feedback coming from one of the social media that is often used by the public, namely Twitter. Tweet or post on the community will be collected based on the time and place specified. Having obtained a collection tweet, would do next text preprocessing stage. Tweet text already passed the stage of preprocessing, for further

processing in sentiment analysis, to determine the positive and negative responses from the public against government policies that have been applied.

Keywords

Text mining, Text preprocessing, Twitter, Nazief and Adriani algorithm, Public Policy

1. Introduction

Social media today have very many users. According id.techinasia.com number of users worldwide in 2015 reached 2.3 billion users. Whereas in Indonesia, the number of active users of social media, reaching 79 million users (Techinasia.com, 2015). According data from Indonesian Ministry of Communication and Information, the population of internet users reached 82 millions users (Kementrian Kominfo, 2015). When examining the trend, the number of users will continue to grow. This is supported by the latest innovations that are always generated by each social media such as Facebook, Twitter, Instagram, Path and others.

A large number of social media users and the users of high growth, leading social media becomes an effective tool to get information from the user. The information referred to as feedback from consumers to a product. Social media can also be used to obtain feedback from the people about policy implemented by the government. One social media which had many users, especially in Indonesia, is Twitter. In the middle of 2015, Twitter had more than 50 million users in Indonesia (Kompas 2016), so it made a large number of posted information that could be taken. Moreover, the characteristics of Twitter are public, informative, real time, and accessisble for any kind of mobile platform.

In this study, it will develop a system to obtain data in the form of feedback coming from Twitter social media that is often used by people. Message from society will be collected based on the time and place specified. Once obtained a collection of messages, will be conducted text preprocessing stage. Messages are already passed the stage of preprocessing text, will be processed in sentiment analysis, to determine the positive and negative responses from the public against government policies that have been applied.

2. Literature Review

2.1 Related works

With the population of blogs and social networks, especially Twitter, opinion mining and sentiment analysis became a field of interest for many researches (Alexander, P., Paroubek, P.

2010). The goal of this research is to develop a data mining system on Twitter with area, time, and other parameter labels. The results of the data mining are used as resources to sentiment analysis, in order to knowing positive and negative resposes of implemented public policy from people. Pradany and Fatichah have discussed about twitter mining to classify sentiment features of Twitter data using K-Medoid Clustering dan Support Vector Machine (SVM) method (Pradany, L. N., Fatichah, C. 2016). In this research, we used Naïve Bayes method to classify opinion data from Twitter. The method is easy and quick to predict the class of the test data set. It also performed well in a multi-class prediction.

2.2 Twitter

Twitter is a social networking service, which allows users to communicate with each other. Users can send and read messages that consist from a maximum of 140 characters, called Tweets (Bryl S., 2014). Twitter contains information that is considered important by users. Messages (tweets) from other users who follow will appear on the home page to read. Users can do a retweet or resend messages sent by other users. In a message sent, when writing the name of another user hence will be written the @ sign followed by a username in the tweet. Users can use the sign # (hashtag) to write messages by topic.

Twitter was founded in March 2006 by Jack Dorsey. Growth in the number of Twitter users rises rapidly, so there are now 500 million Twitter users. The use of Twitter in general has increased dramatically during a popular event, so it appears the Trending Topic feature. Trending Topic is a list of terms that most appear. Twitter rose to second place as a social networking site most visited in the world, from which previously was ranked twenty-two (Bryl S., 2014). The high popularity of Twitter is causing this service has been used for various purposes in various aspects, for example as a tool to convey the aspirations, political campaigns, learning, and as an emergency communication media.

2.3 Twitter Mining

Twitter Mining is one form of mining on social media. The purpose of doing mining on Twitter is that users can obtain data from social media Twitter. Twitter provides a REST API for software developers. REST API is the access program to read and write data on Twitter, like reading tweets, read author profile, knowing the data follower and others. In order to access these data through a REST API, the developer must obtain authority through OAuth. By using OAuth, then developers can safely make a request to the Twitter API (REST API). Response given by the REST API to developers is the data in JSON format. There are some limitations in

the API of Twitter. It depends on the total number of tweets which access via API, but usually you can get tweets for the last 7-8 days (Bryl, S. 2014).

2.4 Sentiment Analysis

Sentiment analysis or opinion mining is the use of natural language processing, text analysis, and computation linguistics, which aims to identify and perform an extraction on subjective information inside an information source, such as web pages, social networks, and other information sources. Sentiment analysis is widely used to obtain a review of the user to an object or a particular topic. In general, the goal of sentiment analysis is to determine the response from users to a particular topic. The response given by the user can be either positive or negative response (Luce L., 2012).

The basis of sentiment analysis is the classification of the polarity from a sentence, text, or document. The classification is done in a simple way, that is with determining a positive, negative or neutral. More complex classification can be used to determine emotion, where the text reflects the emotion of anger, sad, or happy.

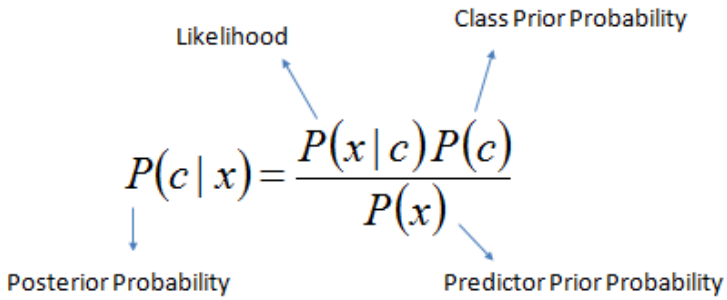
Initially sentiment analysis is used to detect the polarity from a review of a product and film. The polarity detection performed in the document level. In most of the statistical classification method, neutral class ignored on the assumption that a neutral texts is located near the boundary from a binary classifier. Some researchers suggest that, as in every issue of polarity, three categories should be identified. In addition, it can be proven that certain classifiers such as Max Entropy and SVM could benefit from the introduction of neutral class and improve the accuracy from the overall classification. In principle there are two ways to operate with the neutral class. The first way is by identifying the neutral language first, filter it and then assess the rest in terms of positive and negative sentiment. The second way is to build a three-way classification in one step. This second approach often involves estimating a probability distribution over all categories (eg Naive Bayes classifier as implemented by Python NLTK). To use the neutral class depending on the nature of the data: if the data are clearly grouped into neutral language, negative and positive, then it is easier to filter neutral language and focus on the polarity between positive and negative sentiment. Otherwise, most of the data is neutral with a small deviation to the positive and negative so it will be more difficult to clearly distinguish the polarity.

Approach to sentiment analysis can be grouped into three categories, namely knowledge-based techniques, statistical methods, and a hybrid approach. Knowledge-based techniques classify the text by dividing categories based on their clearly emotions in words, such as happy,

sad, scared, angry and bored. Some of the knowledge-based techniques do not only use words that show emotion clearly in the example above, but also include the words associated with the emotion. Statistical methods classify the text by using elements from machine learning, such as latent semantic analysis, support vector machine, "bag of words", and orientation semantics. This statistical method using grammatical relationship of words to obtain opinions, along with the features of the opinion. Grammatical dependency relationship is obtained by parsing deeply. While hybrid approach uses elements from machine learning and knowledge representation, such as ontologies and semantic networks to detect semantic presented in subtle ways, such as through the analysis of concepts that are not explicitly convey relevant information, but that is implicitly associated with other concepts

2.5 Naïve Bayes

Naive Bayes is a classification technique based on Bayes theory assuming independent among predictors. In simple terms, Naive Bayes classifier assumes that the presence of certain features in the class is not associated with the presence of other features. For example, if there are three features that describes an object, then the three features independently contribute to the probability that describes the object, which is called "naive". Naive Bayes model is useful for data sets with a large size.



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Figure 1 : *Naive Bayes Equation (Sayad, Saed 2010)*

Explanation of the equation :

- P (c | x) is the class posterior probabilities (c, target) given predictor (x, attributes).
- P (c) is the class prior probability.
- P (x | c) is a likelihood that indicates the probability of a certain class predictor.
- P (x) is the predictor prior probability.

Naive Bayes models have advantages and disadvantages, which the advantages are:

- It is easy and quick to predict the class of the test data set. It also performed well in a multi-class prediction.
- When there are independent assumptions, a Naive Bayes classifier performs better than the other models such as logistic regression and the need training data less.
- Naive Bayes classifier has performed well in the case of the input category variable compared with numerical variable.

The disadvantages are:

- If a category variable has category that are not observed in the training data set, then the model will assign as "zero probability" and would not be able to make predictions. This is often known as "zero frequency". To solve this, we can use smoothing techniques. The simplest smoothing technique called Laplace estimation.
- On the other side, Naive Bayes also known as bad estimator, so that the probability output is not to be taken too seriously.
- Another limitation is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors that are completely independent.

2.6 NTLK (Natural Language Toolkit)

NLTK is a library suite and program for natural language processing (NLP) which is written in the Python programming language. It was developed by Steven Bird and Edward Lopper at the Department of Computer and Information Science at the University of Pennsylvania. NLTK has a graphic demonstration and sample data. It is accompanied by a book that explains the underlying concept behind the language processing tasks that are supported by toolkit coupled with a cookbook.

NLTK intended to support research and teaching in the field of NLP or closely related fields, including empirical linguistics, cognitive science, artificial intelligence, information retrieval and machine learning. NLTK has been successfully used as a teaching tool, as a means of individual learning, and as a platform for prototyping and building research system. NLTK supports classification, tokenization, stemming, tagging, parsing, and semantic reasoning functions.

3. Implementation and Testing

3.1 Workflow

In this study, we designed a workflow that is shown as in Figure 2. In this figure, explained that:

1. To obtain tweet data, we stream it within 7 days period.
2. Perform preprocessing on the obtained data tweet. After passing the preprocessing stage, tweet data is stored in text files.
3. Texts resulting from the preprocessing stage are still duplicated, therefore duplication tweet is reduced, in order to obtain a unique tweet. The results are stored in text files.
4. The text file that contains the tweet data that has been preprocessed and cleanup duplicate, then manually categorized into two groups: positive and negative tweets. The results are stored in two different text files.
5. Both the text file used as a training set for the machine. The training set as the reference in determining tweet polarity that will be tested.

Figure 2 : *The Workflow*

3.2 Twitter Authentication

Twitter providing facilities for the software developer to obtain data on Twitter via the REST API or known by Twitter API. This facility enables software developers to retrieve data such as tweet content, location and place, writer and follower from tweet writer. Twitter API can only be used if the software developer is authenticated in the Twitter application. In order to be authenticated in the Twitter, the developer must register themselves and the applications. After registering, the software developers will get some form of key alphanumeric code, which will be written in the software developed.

In order to authenticated, software developers must sign to https://apps.twitter.com/ address. If the developer has never made a previous application, then they should make a new application by selecting Create New App button, and fill out the form provided. If the software developers already have an application, then the developers just choose the name of his

application that appears on the screen. The key that will be used to build applications that are in Keys tab and Access Token.

3.3 Implementation

The making of Twitter sentiment analysis aims to determine the tweet sent by the public is positive or negative in response to government policy. By knowing the response from the public, the government is expected to act on the policy. To determine its positive or negative tweets, it is necessary to test the polarity. Flow of determining the polarity test can be seen in Figure 3.

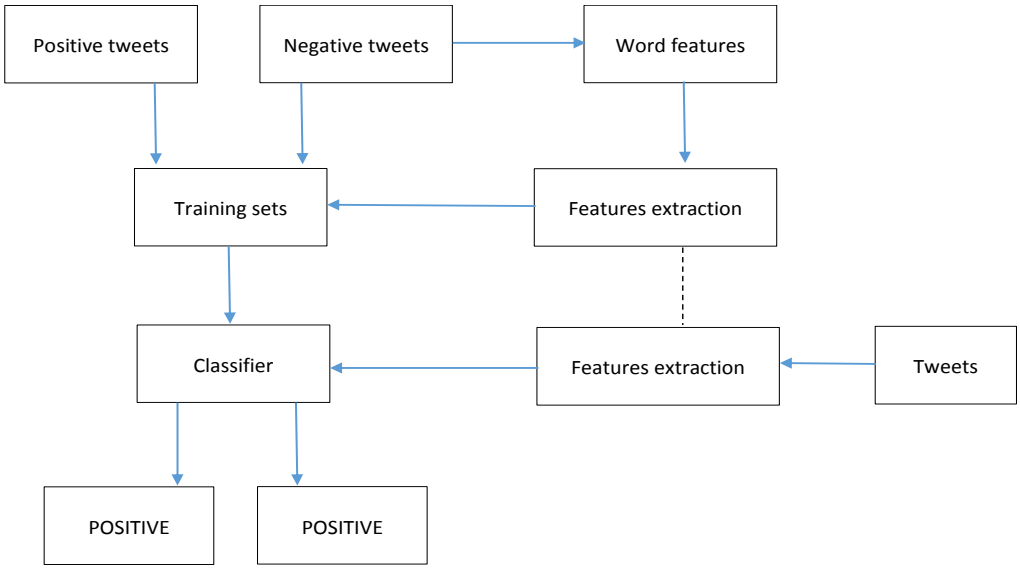


Figure 3 : *Polarity Test Flow*

The Naive Bayes classifier uses the prior probability of each label which is the frequency of each label in the training set, and the contribution from each feature. In our case, the frequency of each label is the same for 'positive' and 'negative' (Luce, L. 2012). Before performing polarity test, tweet data must be obtained first by streaming. After the tweet data is obtained, then performed the preprocessing and continued with grouping into a positive and negative tweets. Then we can perform a polarity test against the new tweet data. The process is described in Figure 2. For the polarity test itself is described in Figure 3. For the detail section to the results obtained in the form of positive and negative determination is described in the next section.

3.3.1 Streaming Tweet

To collect large amounts tweet simultaneously by keyword, location, and specific time can use streaming techniques. There are two ways to streaming data, that crawling and scraping. For this study using crawling techniques. Crawling techniques implementation in this study using R language. Before crawling, it must insert four key first in order to access the Twitter API,

which is api_key, api_secret, access_token, and access_token_secret. By using searchTwitter function, then tweet with keywords can be searched in a certain amount.

3.3.2 Preprocessing

Stages of preprocessing that is performed include:

1. Eliminate Retweet entities, which is RT or via strings
2. Eliminate the usernames on Twitter, username is marked with an @ sign, followed by its name
3. Eliminate punctuations such as . , ! ?
4. Eliminate numbers
5. Eliminate HTML links, like: http://domainname.com/x.html
6. Eliminate unnecessary spaces

After passing the preprocessing stage, tweet data is stored in text files.

3.3.3 Remove Duplicate Tweet

The text file generated from the preprocessing stage are still have a double tweets. Double tweets will affect the calculation results of occurrence frequency of a word, so to improve the calculation results, then it should perform the process to remove double tweets. To remove double tweets, in this study used the code in the Python language.

3.3.4 Manual Labelling

The result of removing duplicate tweets is a text file. Then the text file labeled positive or negative manually. To simplify this process, we can use the MS Excel application. The labels consist of "post" and "neg". The "Pos" label for positive tweets, and the "neg" label for negative tweets like in the picture.

After all tweets are labeled, then it's all divided into a set of positive tweets and a set of negative tweets. Then a set of positive tweets copied on a separate text file, as well as to negative tweets. So from this process obtained two text files, one of which contains a set of positive tweets, and the other contains a set of negative tweets.

3.3.5 Polarity Test

Polarity test is used to test a new tweets, and determine positive or negative tweets. Flow to determine the type of tweets is illustrated in Figure 3. Explanation of the process is as follows:

1. Extracting word features from a text file that contains a set of positive and negative tweets
2. The results of features and tweets extraction is used to training set.
3. Classify the training set using a Naïve Bayes Classifier function owned by NLTK.
4. Extract the features of new tweets.

5. Classify the new tweets based on training set that has been classified.
6. The new tweets are classified into positive and negative.

The results of the program code when executed are as follows:

```
Most Informative Features
         siapkan = True     positi : negati =     9.7 : 1.0
           paket = True     positi : negati =     7.2 : 1.0
        keluarkan = True     positi : negati =     6.3 : 1.0
          ekonomi = True     positi : negati =     6.2 : 1.0
          membuat = True     negati : positi =     5.7 : 1.0
       kepedulian = True     positi : negati =     5.0 : 1.0
         kesehatan = True     positi : negati =     5.0 : 1.0
         pengusaha = True     negati : positi =     5.0 : 1.0
            petani = True     negati : positi =     5.0 : 1.0
             hukum = True     positi : negati =     4.9 : 1.0
{'negative': ['Pengusaha tidak memberikan libur bagi karyawannya',
'Tidak  berpihak  pada  petani'],  'positive':  ['Itu  merupakan
kepedulian hukum']}

Process finished with exit code 0
```

Figure 4 : *The result of program*

3.4 Testing

To determine technical performance used in the application are counting accuracy, precission and recall. Precision measures the exactness of a classifier. A higher precision means less false positives, while a lower precision means more false positives. This is often at odds with recall, as an easy way to improve precision is to decrease recall (Perkins, J. 2010). In this calculation, we used a text file that contains the positive and negative tweets. Below is an explanation of the calculation:

1. Extracting word features from text file that contains a set of positive and negative tweets.
2. Took ¾ of fetures from each set of positive and negative tweets.
3. To produce training data, ¾ parts of positive tweet features added by ¾ parts of negative tweet features.
4. To produce data test, ¼ parts of positive tweet features added by ¼ parts of negative tweet features.
5. Classify training data using NaiveBayesClassifier function provided by NLTK
6. Test data is classified based on training data
7. Calculating the accuracy, precission, and recall using accuracy, precission, and recall function provided by NTLK

The calculation of accuracy, precission, and recall results as follow:

```
using all words as features
len of positive features 150
len of negative features 150
pos_cutoff 112 neg_cutoff 112
train on 224 instances, test on 76 instances
accuracy: 0.6578947368421053
pos precision: 0.75
pos recall: 0.47368421052631576
neg precision: 0.6153846153846154
neg recall: 0.8421052631578947
Most Informative Features
                  paket = True      pos : neg    =     11.4 : 1.0
                siapkan = True      pos : neg    =      8.3 : 1.0
                ekonomi = True      pos : neg    =      8.3 : 1.0
                  hukum = True      pos : neg    =      7.4 : 1.0
                   jadi = True      neg : pos    =      6.3 : 1.0
                membuat = True      neg : pos    =      5.7 : 1.0
                 petani = True      neg : pos    =      5.0 : 1.0
                  tidak = True      neg : pos    =      5.0 : 1.0
                 sektor = True      pos : neg    =      4.3 : 1.0
              kepedulian = True      pos : neg    =      4.3 : 1.0

Process finished with exit code 0
```

Figure 5 : *The calculation of accuracy, precission, and recall*

Getting the results:

- Accuracy : 66%
- Positive tweets accuration : 75%
- Positive tweets recall : 47%
- Negative tweets accuration : 61%
- Negative tweets recall : 84%

4. Conclusion

Based on the results of testing to perform sentiment analysis on community feedback, we conclude that:

1. The combination of search keywords on Twitter, especially when using Indonesian Language, greatly affect the outcome of tweets that appear to do the crawling. Flagging # followed by space also affects the results tweet.

2. If using the Twitter API and using crawling techniques, then the tweets obtained limited to tweets that was sent within the previous 7 days.
3. Tweet in Indonesian has a non-standard words, eg slang words and abbreviations. So it requires more complex preprocessing.
4. The retweet causing double tweet in the collection, so it requires a double tweets cleaning to eliminate them.
5. With the number of tweets in the form of news sentences have lead any difficulties in determining the polarity manually. It is caused by most of the news was neutral sentences of Indonesian Language.
6. In a supervised learning method, the input data for training greatly affect system performance. If the training data is ideal, then result good outcomes.

REFERENCES

Alexander, P., Paroubek, P. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Proceedings of the Seventh conference on International Language Resources and Evaluation (1320-1326).

Bryl, S. 2014. Twitter Sentiment Analysis With R. Diperoleh 30 September 2016. Retrieved from http://analyzecore.com/2014/04/28/twitter-sentiment-analysis.

Cho, Hyun-Woong and Kim, Woo Je. 2015. Development of a k-nn model to predict the Polarity of korean game review comments. Retrieved from https://grdspublishing.org/download.php?table=MATTER%20%20%20&id=MSV1I1108119. MATTER: International Journal of Science and Technology. Special Issue Vol.1 Issue 1, pp. 108-119.

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. 2008. Introduction to Information Retrieval. Retrievevd from http://www-nlp.stanford.edu/IR-book/. Cambridge University Press.

Kementerian Kominfo. 2016. Internet users in Indonesia. Retrieved from https://kominfo.go.id/index.php/content/detail/3980/Kemkominfo%3A+Pengguna+Internet+di+Indonesia+Capai+82+Juta/0/berita_satker

Kompas.com. 2015. Pengguna Twitter di Indonesia Capai 50 Juta. Retrieved from http://tekno.kompas.com/read/2015/03/26/16465417/Pengguna.Twitter.di.Indonesia.Capai.50.Juta

Luce, L. 2012. Twitter sentiment analysis using Python and NLTK. Retrieved from

http://www.laurentluce.com/posts/twitter-sentiment-analysis-using-python-and-nltk/

Perkins, J. 2010. Text Classification For Sentiment Analysis – Precision And Recall. Retrieved from http://streamhacker.com/2010/05/17/text-classification-sentiment-analysis-precision- recall/

Pradany, L. N., Fatichah, C. 2016. Analisa Sentimen Kebijakan Pemerintah Pada Konten Twitter Berbahasa Indonesia Menggunakan SVM dan K-Medoid Clustering. SCAN Vol. XI (59-66) No. 1 ISSN : 1978-0087, Februari 2016

Sayad, Saed. 2010. Data Mining Classification with Naïve Bayesian. Retrieved from http://www.saedsayad.com/naive_bayesian.htm

Techinasia.com, 2015, Statistik Pengguna Internet dan Media Sosial Terbaru di Indonesia. Retrieved from https://id.techinasia.com/talk/statistik-pengguna-internet-dan-media-sosial-terbaru-di-indonesia