Journal of Information System, Applied, Management, Accounting and Research
JISAMAR
ISSN: 2598-8719 (Online)
ISSN: 2598-8700 (Printed)
Vol. 2 No.2 Mei 2018

SENTIMENT ANALYSIS ON TWITTER ACCOUNT USING NAIVE BAYES CLASSIFIER ALGORITHM

Case Study: Indonesia Healthcare and Social Security Agency (BPJS Kesehatan)

Silvi Dwi Andrianti
BPS-Statistics Indonesia
Jakarta-Indonesia
silvidwi@bps.go.id

Abstract

BPJS Kesehatan is organizing the health care insurance for all Indonesian people. By 2018, the number of participants of BPJS Kesehatan reached 196,662,064 people. A large number of these users make BPJS Kesehatan must provide services in the form of feedback. One uses media is Twitter. Information obtained from any tweets, can be used as a tool of policy makers and this can be done by using sentiment analysis. At sentiment analysis, a classification method that can be used is Naive Bayes classifier algorithm. Naive Bayes classifier algorithm is a classification method that is rooted in the Bayes theorem. In this paper we show a system of sentiment analysis BPJS twitter account with a Naive Bayes classifier algorithm. Naive Bayes classifier algorithm consists of two stages. The first stage is to set the sample document training (training data) and the second stage is the process of classifying documents of unknown category (class). The system uses a method Naive Bayes classifier algorithm for classification. Phase to be conducted before the classification is preprocessing. Stages in the preprocessing consists of a folding case, normalization features, the emoticons convert, convert negation, tokenizing, stemming and stopword removal. Tweets that have passed the stage of preprocessing will be classified into positive opinion or negative opinion and displayed in a pie chart. Based on testing, the results tweets classification accuracy is 88% with precision positive 85%, precision negative 75% and precision neutral 92%.

Keywords— BPJS Kesehatan; Twitter; Sentiment Analysis; Naive Bayes Classifier Algorithm

I. INTRODUCTION

BPJS Kesehatan (Healthcare and Social Security Agency) is a State-Owned Enterprises were specially commissioned by the government to administer health care benefits for all Indonesian people, especially for Civil Servants, Pension Recipients civil servants and TNI / POLICE, Veterans, Independence Pioneers and their families and other business entities or commoner. By 2018, the number of participants BPJS Kesehatan reached 196,662,064 people [1]. A large number of participants makes BPJS must provide services in the form of feedback to participants in order to facilitate the review of services. One of the media is used as a feedback service is Twitter.

At the present time, microblogging became an online communications tool that is very popular in the community. Looks very clear on one very popular microblogging is Twitter,the number of active users reached more than 645.7 million with the average number of tweets per day to 58 million tweets [2]. Indonesia is the third largest Twitter users in the world, with the amount of 6.5 percent below the United States (24.3 percent), and Japan (9.3 percent) [3]. Number of tweets is expanding rapidly due to its simplicity and ease of use are some of the reasons why Twitter is more popular with the people of Indonesia in communicating. Of course, the information contained in these tweets is very valuable as a tool of policy makers and this can be done by using sentiment analysis.

Sentiment analysis or opinion mining is also called, is a field of study in analyzing people's opinion, evaluation, assessment, attitudes and emotions to an entity such as products, services, organizations, individuals, issues, events and topics. The main focus of the analysis is to express sentiments which includes a positive opinion, and which includes a negative opinion [4]. One example of the application of sentiment analysis is when a company issues a product and a company that provides services to receive the opinions of consumers regarding these products. Sentiment analysis is used to classify the opinions of the positive and negative of consumers who use the product so as to speed up and simplify the task of companies to revisit the shortcomings of its products.

Based on the picture above explanation, here the authors are interested in applying sentiment analysis

Journal of Information System, Applied, Management, Accounting and Research
JISAMAR
ISSN: 2598-8719 (Online)
ISSN: 2598-8700 (Printed)
Vol. 2 No.2 Mei 2018

system to the official Twitter account BPJS Kesehatan using Naïve Bayes Classifier (NBC) algorithm. The purpose of this research is to apply the methods Naïve Bayes Classifier (NBC) algorithm in building an application that is able to classify it as an opinion positive opinion or negative opinion as one of the main functions of sentiment analysis.

The objectives to be achieved in this research is to build a system of sentiment analysis on the official Twitter account BPJS Kesehatan using Naïve Bayes Classifier (NBC) algorithm, which can help in analyzing sentiment BPJS users. In addition to determining that a trigger word sentiment based tweets every day.

II. RELATED WORK

Research on sentiment analysis using social media has been a lot done. Some research about social media analysis that has been done is as follows.

In [5] the author analyze Twitter postings of 140 characters or less, known as tweets, to infer user health status over time.

In [6] show that a SVM classifier combined with a cluster ensemble can offer better classification accuracies than a stand-alone SVM. In our study, we employed an algorithm, named C3E-SL, capable to combine classifier and cluster ensembles. This algorithm can refine tweet classifications from additional information provided by clusters, assuming that similar instances from the same clusters are more likely to share the same class label.

In [7] the author present how sentiment analysis can assist language learning, by stimulating the educational process and experimental results on the Naive Bayes Classifier.

In [8] explained the detailed work done in developing a system which can be used for the purpose of opinion analysis of a product or a service. The system readily processes the tweets by pulling data from tweeter posts, preprocessing it and connecting to Alchemy API by REST call method and showing the result graphically

III. METHOD

In this study, a method to build sentiment analysis system will pass through the stages in Figure 1:

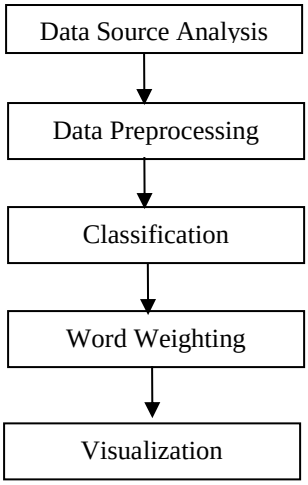


Fig 1. Method to build sentiment analysis system

A. Data Source

The data used in this study was taken from a collection of tweets Indonesian taken from the official Twitter account BPJS Kesehatan. The data is obtained by making tweets crawling program that uses the Tweetinvi API. In the process of crawling, automatically retrieve data tweets that contain the word "BPJSKesehatanRI". Tweets collected data will pass through the preprocessing stage and will be classified. In this sentiment analysis system, tweets will be classified into three classes (categories), namely the class of positive sentiment, sentiment is neutral and negative sentiment class.

Fig 2. Example of tweet contain "BPJSKesehatanRI"

B. Data Preprocessing

Text processing is the process of digging, process, organize the information by analyzing the relationship, the rules that exist in the textual data semi-structured or unstructured. To be more effective in the process of processing carried out step transformation of data into a format that makes it easy

Journal of Information System, Applied, Management, Accounting and Research
ISSN: 2598-8719 (Online)
ISSN: 2598-8700 (Printed)
Vol. 2 No.2 Mei 2018

for the user's needs. Preprocessing is one important step in sentiment analysis. Similarly preprocessing on Information Retrieval (IR), stage consists of tokenizing, normalization features, case folding, stop-word removal and stemming. But the sentiment analysis preprocessing, there are additional stages such as emoticons convert and negation convert.

C. Naïve Bayes Classifier (NBC) algorithm

Naïve Bayes Classifier (NBC) algorithm is a classification method that is rooted in the Bayes theorem. The main characteristic of the Naïve Bayes classifier is a very strong assumption (naive) will be independent of each variable. In other words, Naïve Bayes classifier assumes that the existence of an attribute (variable) has nothing to do with the existence of attributes (variables) to another. Naïve Bayes classifier algorithm consists of two stages. The first stage is to set the sample document training (training data) and the second stage is the process of classifying documents of unknown category (class).

This algorithm utilizes probability theory put forward by the British scientist Thomas Bayes, that predict the probability in the future based on past experience. Because assumptions unrelated attributes (conditionally independent), then:

$$Vmap = argmax\ Vj \in V\ P(Vj)\ \Pi\ P(wk \mid Vj) \quad (1)$$

Having obtained the calculations for each category, the selected category is one that has the greatest Vmap value. The value of P (Vj) is determined at the time of training, whose value is based on the equation:

$$P(Vj) = |docs\ j| / |example| \quad (2)$$

Where P (Vj) is the probability of each document to a collection of documents. | docs j | is the number of documents that have the category j in training. | example | is the number of documents in the examples used during the training.

For the value of P (wk | Vj) is determined by the equation:

$$P(wk \mid Vj) = |nk+1| / (n+|kosakata|) \quad (3)$$

Where P (nk | Vj) is probability of occurrence of the word wk in a document by category Vj. Nk is the frequency of appearance of the word in the document that categorized wk Vj. n for number of all the words in the document category Vj. | $kosakata$ | is the number of words in the training examples.

D. TF-IDF

In sentiment analysis, weighting the word is used to obtain a topic or keyword from the collection of sentiment. One weighting method is TF-IDF (Term Frequency - Inverse Document Frequency). The value weight of a word (term) expressed interest in representing the weight of tweets. On the TF-IDF weighting, the weights will be even greater if the higher frequency of occurrence of the word, but the weight will be reduced if the word is more often appears in other tweets.

TF-IDF method is a method of weighting in the form of a method which is the integration between the term frequency (TF), and the inverse document frequency (IDF). TF-IDF method can be formulated as follows:

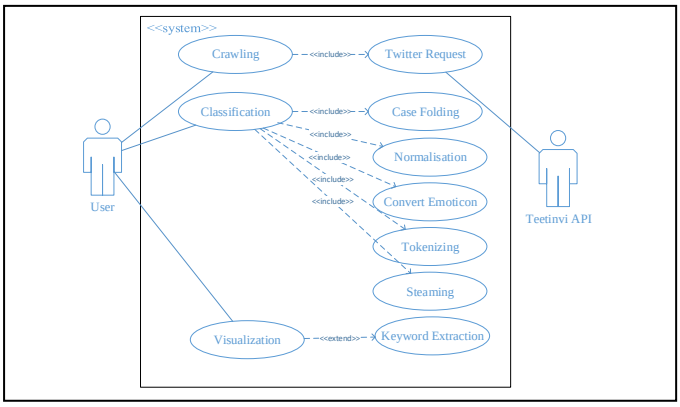
$$tf.idf(d,t) = tf(d,t) * \log(|N|/df) \quad (4)$$

Where tf (d, t) is the frequency of occurrence of the word t in the document d. | N | is the sum of all documents in the collection, and df is the number of documents that contain the word t. TF-IDF weighting method used for weighting method is most excellent in information retrieval task. Weight value of a term expressed interest in representing the weight of the document. On the TF-IDF weighting, the weights will be even greater if the frequency of occurrence of the higher term, but the weight will be reduced if the term is more often appears in other documents.

IV. DESIGN AND IMPLEMENTATION

A. Use Case Diagram

Use case diagram shows the relationships that occur between actors with the use case in the system. Use case is designed to help prospective users of the system to get a full understanding of the system to be built. Use case diagrams sentiment analysis can be seen in Figure 3 below.



Journal of Information System, Applied, Management, Accounting and Research
ISSN: 2598-8719 (Online)
ISSN: 2598-8700 (Printed)
Vol. 2 No.2 Mei 2018

Fig 3. Use case diagram sentiment analysis

B. Activity Diagrams

Activity diagrams in this study describes the workflow stages of activity use case to be built. Here are each activity diagram sentiment analysis.

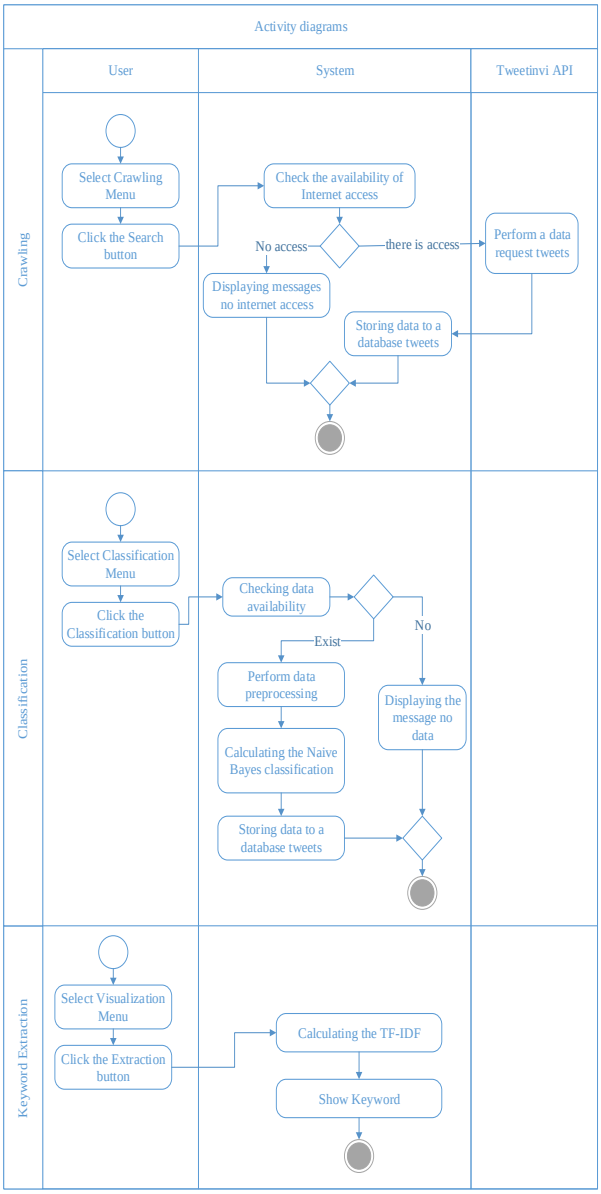


Fig 4. Activity diagram sentiment analysis

C. Implementation

Implementation is the stage where the system is ready to operate on a real stage, so it will be known whether the system has been made completely as planned. In the implementation of this software will be explained how this system works, to give the appearance of a system or application is made.

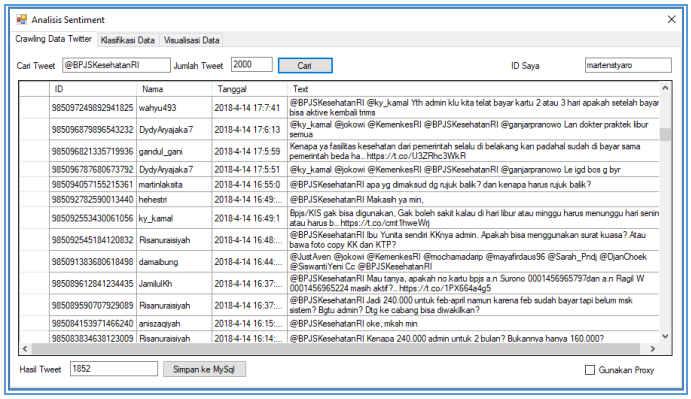


Fig 5. Crawling Menu

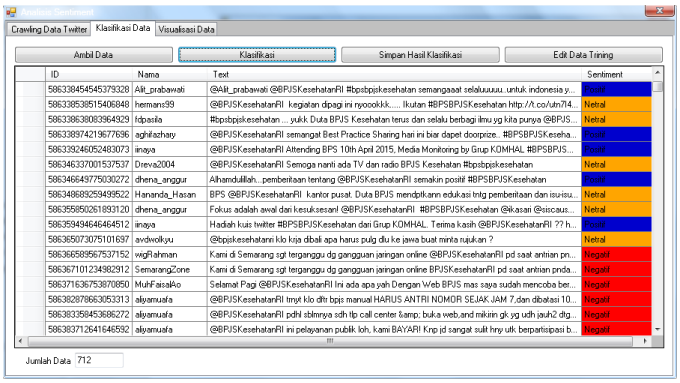


Fig 6. Classification Menu

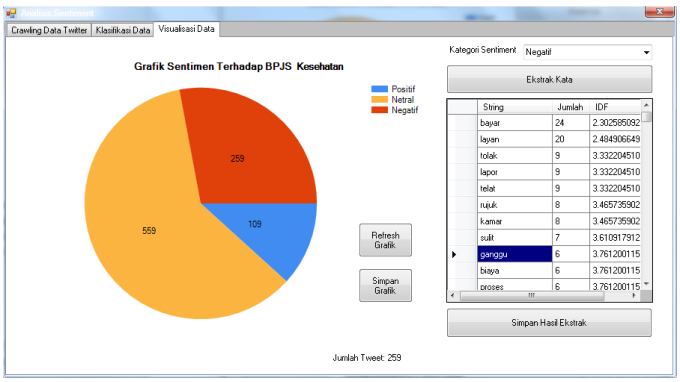


Fig 7. Visualization Menu

Journal of Information System, Applied, Management, Accounting and Research
JISAMAR
ISSN: 2598-8719 (Online)
ISSN: 2598-8700 (Printed)
Vol. 2 No.2 Mei 2018

V. EXPERIMENTAL RESULT

Tweets classification accuracy testing is performed to determine the level of classification accuracy tweets which performed manually with tweets classification done by the system by using naïve Bayes classifier algorithm. Tests performed by using the confusion matrix that is a matrix of predictions will be compared with the original class of the input data. Tests performed using the data 380 tweets which taken randomly and are already labeled. Tweets Data will be compared with the results of the classification performed by the system. Results of testing the accuracy of classification can be seen in the following tweets.

TABLE I. CONFUSION MATRIX

Actual Class	**Predicted Class**		
	Positive	***Negative***	***Neutral***
Positive	11	0	2
Negative	3	74	21
Neutral	12	9	248

Accuracy = (11+74+248)/380 = 0.88
Precision Positive =11/13 = 0.85
Precision Negative =74/98 = 0.75
Precision Neutral =248/269= 0.92

The accuracy of the test data used in Table 1 as many as 380 tweets, which consists of 26 tweets positive, 83 tweets negative and 271 tweets neutral. The results of classification done by the system, as many as 13 tweets including positive sentiment, 98 tweets including negative sentiment, and 269 tweets including neutral sentiment, then the number of correct classification is 28 tweets. Based on testing accuracy, the results obtained classification accuracy tweets from sentiment analysis system using a naïve Bayes classifier algorithm at 88%. With precision was equivalent to 85% positive, 75% of negative precision and 92% of neutral precision. The conclusion from this is that the testing accuracy of naïve Bayes classifier algorithm can be used as a method of classifying the sentiment analysis for a large degree of accuracy.

Extraction results from trigger word for the analysis can be seen in the following table:

TABLE II. EXTRACTION WORD

Positive			***Negative***		
Word	***Total***	***IDF***	***Word***	***Totale***	***IDF***
layan	6	2.39	telat	8	2.83
sehat	6	2.39	tolak	7	2.94
bantu	5	2.56	kamar	7	2.94
e_senang	5	2.56	sulit	7	2.94
mudah	5	2.56	rujuk	6	3.091
selamat	5	2.56	ganggu	6	3.091
proses	4	2.83	urus	6	3.091
terimakasih	4	2.83	tanggap	6	3.091
gratis	4	2.83	coba	5	3.29
laksana	3	3.13	gagal	5	3.29

VI. CONCLUSION

Based on the results of the implementation and testing has been done on sentiment analysis system using Naïve Bayes classifier method is it can be concluded that the method Naïve Bayes Classifier (NBC) algorithm can be used to classify tweets on sentiment analysis system. This sentiment analysis systems provide information on the percentage of positive and negative sentiment depicted in the form of pie charts and information about the words that affect the sentiment word.

Suggestions for further development, as follows: The addition of features to sort out tweets that include non-opinion or opinions. Increasing the number of training data to get a better result when the classification tweets. The addition of features to address the imbalance training data, in order to obtain optimal results when the classification process tweets.

REFERENCES

[1] BPJS Kesehatan, “Jumlah Peserta” 2018. [Online]. Available: http://bpjs-kesehatan.go.id/bpjs/index.php/home

[2] "statisticbrain," [Online]. Available on : http://www.statisticbrain.com/twitter-statistics/. [Accessed 23 March 2018].

[3] Arifin, “enciety.co”, 10 Februari 2018. [Online]. Available: http://www.enciety.co/pengguna-twitter-indonesia-terbanyak-ketiga-dunia/ [Accessed 23 March 2018]

[4] B. Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publisher, 2012.

[5] Kashyap, R., Nahapetian, A., “Tweet analysis for user health monitoring”, *Wireless Mobile Communication and Healthcare (Mobihealth), IEEE Conference Publications*, pp: 348 - 351, 2014

[6] Coletta, L.F.S., da Silva, N.F.F., Hruschka, E.R., Hruschka, E.R., “Combining Classification and Clustering for Tweet Sentiment Analysis”, *Intelligent Systems (BRACIS), 2014 Brazilian Conference, IEEE Conference Publications*, pp: 210 – 215, 2014

[7] Troussas, C., Virvou, M., Junshean Espinosa, K., Llaguno, K., Caro, J. “Sentiment analysis of Facebook statuses using Naive Bayes classifier forlanguage learning”, *Information,*

ISSN: 2598-8719 (Online)
ISSN: 2598-8700 (Printed)
Vol. 2 No.2 Mei 2018

Intelligence, Systems and Applications (IISA), 2013 Fourth International Conference, IEEE Conference Publications, pp: 1 – 6, 2013

[8] Das, T.K., Acharjya, D.P., Patra, M.R., "Opinion mining about a product by analyzing public tweets in Twitter", *Computer Communication and Informatics (ICCCI), 2014 International Conference, IEEE Conference Publications*, pp: 1 – 4. 2014