IJER Indonesian Journal of Educational Review
p-ISSN 2338-2018 | e-ISSN 2335-8407
Vol. 5, No.2 , Desember 2018, p

Available online at
http://pps.unj.ac.id/journal/ijer

BIG DATA FOR CLASSIFICATION OF COMMUNITY COMPLAINTS AGAINST PUBLIC SERVICES ON TWITTER

Muqorobin[1], Zul Hisyam[2], Arief Setyanto[3]
AMIKOM Yogyakarta University
robbyaullah@gmail.com[1]
zul0342@gmail.com[2]
arief_s@amikom.ac.id[3]

Abstract

....

....

The development of the internet today is very rapid and has a great influence on human life, one of which is the increasingly rapid dissemination of information. At present almost all information that we want is on the internet, both from websites, blogs, social media, and others. One of the most popular media in sharing information is Twitter social media. Twitter is a social network that allows each user to share information [1]. Besides being used as social media, Twitter is also used as a medium to read news. Users who want to get information from an account must first become a follower of the account. The large amount of information available on Twitter can also be used to conduct a study, especially in terms of sentiment analysis by categorizing information based on certain categories.

Text categorization methods that can be used at this time are quite large, including Bayes classification algorithm, K-Nearest Neighbor (KNN), Neural Network (NN), Support Vector Machine (SVM), The Decision Tree, K-Means, etc.

One method that can be used is Naive Bayes. This approach is an approach that refers to the Bayes theorem which is a statistical principle of opportunity to combine previous knowledge with new knowledge, and then this principle is used to solve classification problems [2].

In a previous study conducted by Veronikha Effendy, et al. (2016), the SVM method was used to determine positive and negative opinions on public transportation via twitter, which then the results of the analysis will be used to determine the factors that are the main causes of the inability to use public transportation and factors that make people choose to use this type of transportation[3].

Based on the research conducted above the reason the authors conducted this research is to be able to predict by classifying complaints or not on public services based on 3 parameters, Transportation, Hospital, and Market. The results of this analysis can be used to assist in predicting the occurrence of complaints or not in the future. So that with this information will be able to help the government in deciding general policies.

METHOD

Naïve Bayes

Naïve Bayes algorithm is one type of classification method with probability and statistical methods. This method was put forward by British scientist Thomas Bayes, namely predicting opportunities in the future based on previous experience so that it is known as the Bayes theorem. The theorem is assumed to have mutually independent attributes[4].

The theoretical basis used in carrying out this classification process is the Bayes theorem shown by the equation (1).

1. P (A|B) = (p (B|A)* p(A)) / p(B)
   Explanation :
   P(A|B) : Posterior A value when B
   p(B|A) : B Likelihood value when A
   p(A) : Prior Value in class A
   p(B) : Evidence value of a class

At opportunity A as B, it is obtained from opportunity B when A multiplied by opportunity A and divided by opportunity B. The use of Naïve Bayes on a data with more than one feature / attribute causes equation (1) to be more complex as shown in equation (2).

2. $$P(A|B_l \ldots\ldots B_n) = \frac{P(A)P(B_1 \ldots\ldots B_n|A)}{p(B_1 \ldots\ldots B_n)}$$

The value of P (B1 ... Bn) is constant for each experiment so that the maximum value of a class is determined by the maximum value between P (A) P (B1 …… Bn | A). The function of the equation formed into a maximum multiplication for the prior value and the likelihood function, the function is shown in the equation (3).

3. $$f_c(F) = \underset{c \in F}{\arg max} P(A)\left(\prod_{i=1}^{n} P(B_i|A)\right)$$

Clasification Twitter System Experiments

The System Experiment conducted aims to create an analysis system in classifying tweets using the Naïve Bayes method. At this stage it is known that the suitability of the output produced by the system and the most suitable parameters to be used as input in the classification process. The analysis in this study is to classify tweets using Indonesian[5].

The following are the stages in making a system :

1. Data collection of tweets from each account obtained through the twitter API.
2. Pre-processing, which is cleaning noise such as HTML tags, username, hashtag, url, link. Then from the data obtained, the word normalization process is carried out. The process of word normalization, changing the word abbreviation or non-standard word or slang into a word in standard form according to the large Indonesian dictionary [6].
3. Tokenisation, which is to change the sentence into a form of tokens using a delimiter or space bar. The token used is unigram (consisting of one word).
4. Stemming, which is taking the basic word in a tweet by removing the prefix, suffix or both.
5. Feature extraction, which is observing the frequency distribution of word appearance and number of features. the best threshold value is the point where the frequency of occurrence of words and features start to be constant.
6. Classification of training data using the Naïve Bayes method. Training data contains files of the attributes specified according to the parameters used (total of transportation complaints, total of market complaints, total of hospital complaints). The result of the classification is the content of complaints from each tweet on the parameters: transportation, hospital and market on each data in the training data.

Analisis Clasification

Analysis in this research the author will process the data sourced from the Twitter post. The purpose of this study is to predict complaints or not on public services based on activities in posts on Twitter. System built by applying the Naïve Bayes method. At the complaint classification stage using training data from the results of previous experiments. Data processing in this study will use tweet data of 700 tweets or records. From 700 data records, the researcher will divide into two parts, 75% for training data, 525 records and 25% for data testing, which is 175 records. Because in determining the good proportion for training data is 75% and testing data is 25%.

Performance Evaluation

System testing is useful to determine the level of accuracy of system performance. In Big Data for the classification of public service complaints there are several parameters that are mutually influential in predicting public service complaints. For this reason, in evaluating the performance of classifiers, the researchers used the Confusion Matrix test with Accuracy, Precision, Recall, F-Measure tests on the system that had been built. The system testing formula with the matrix confusion method is shown in the equation (4).

4.
$$Akurasi=\frac{\sum_{i=1}^{l}\frac{TP_i+TN_i}{TP_i+TN_i+FP_i+FN_i}}{l}*100\%$$

$$Presisi=\frac{\sum_{i=1}^{l}TP_i}{\sum_{i=1}^{l}(FP_i+TP_i)}*100\%$$

$$Recall=\frac{\sum_{i=1}^{l}TP_i}{\sum_{i=1}^{l}(TP_i+FN_i)}*100\%$$

$$F-measure=\frac{2*precision*recall}{precision+recall}*100\%$$

The results of the Accuracy, Recall, Precision, F-Measure values produce a percentage value from 0% - 100% and good results are values above 70% and close to 100%[7].

RESULT AND DISCUSSION

Big Data in the classification of public service complaints is expected to make future predictions on public service complaints. The system was built by taking the twitter post data of the Indonesian community. From the results of twitter posts that have been post by the community, it generates large data collection so that the implementation of machine learning can help a model in classification of exclusion and not complaints on public services.[8].

Data Analysis

The data used in this study comes from geolocation posts in the Indonesian region. Retrieving data using R studio tools by taking word categories based on parameters, namely transportation, hospital and market [9]. The results of data retrieval are still in the form of raw data that does not yet have a label, so before conducting the machine learning process the researcher conducts the cleaning data and labelling process, so that it becomes a dataset that can be used in predicting public service complaints. Retrieval of initial data as many as 1000 data, then after carrying out the cleaning process data obtained 700 data. The data is then labelled based on the types of parameters such as: transportation, hospitals and markets. For the results of class classification namely Complaints and Not Complaints.

Text Processing

Text pre-processing is a set of steps that must be done to prepare a collection of datasets into the input data in the next process, namely classification using Naïve Bayes. The several steps carried out on this text pre-processing are tokenizing, stopword removal, and stemming [10].

Tokenizing process is the process of separating each word in a sentence so as to produce a collection of independent words. Separation of words is done by finding space between words [11]. In this process, punctuation is also done. The next step is to do the filtering process. In this process, each word that has stood alone will be identified to determine which word will be used or deleted. Deleted words are words included in the stoplist. A stoplist is non-descriptive words that can be discarded in the bag-of-words approach. Examples of stopwords are and, at, which, from, or, at, during, and so on. In this study focusing on the form of complaints from the tweet text, the words that contain the meaning of the tweet entity such as mention, retweet, hashtag, and url links will also be deleted. Each word will also be cleared of symbols or noisy text, such as: (_~&#([0-9]+);‘) [12]. The several stages carried out in the stemming process in detail are as follows :

1. Word checking by adjusting the standard words in the Indonesian language dictionary.
2. Remove additional words or inflection suffixes, namely: "-lah", "-kah", "-ku", "- mu", or "it"
3. Check prefixes and suffixes that are not allowed, namely: ("be-" and "-i"), ("di-" and "-an"), ("ke" and "-i, -kan"), ("me" and "-an"), ("se" and "-i, -kan")
4. Remove derivation suffixes and prefixes that are: "-i", "-an", "di-", "to", "se", "te-", "be-", "me", "pe"

Training dan Testing

Datasets that already have labels based on parameter and class attributes, can be used to process machine learning using the Naïve Bayes method. Based on the results of grouping data that has been done by researchers, it can be seen in table 1.

Table 1. Parameter Data

No	Category	Complaint Class			Not Complaint Class		
		Transportation	Hospital	Market	Transportation	Hospital	Market
1	Yes	170	124	38	197	119	184
2	No	128	174	260	205	283	218

The parameter data in table 1 is useful for explaining the amount of data in the parameters included in the Yes and No categories. For the Yes category it means that the parameter has an influence on class data while for the No category it means that the parameter has no effect on class data. For more details, it can be described in Figure 1. Views that include the Yes and Figure 2. Views that are included in the category No.

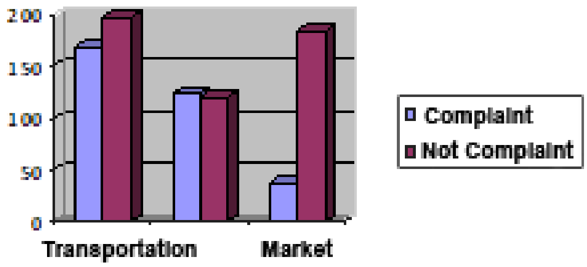

Fig. 1. Category Yes in Parameters

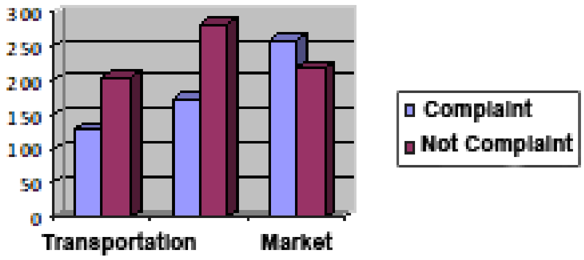

Fig. 2. Category No in Parameters

Based on the results of comparisons in Figures 1 and 2, it can be concluded that the posts that discuss general service availability are quite small. Because the number of class grades that are not complaints is more than the amount of the complaint. This shows that the Indonesian people are quite satisfied with the public services provided by the government. To measure the system, an accuracy test will be conducted using the matrix confusion method.

Testing Result

In implementing data processing researchers use data as many as 700 data records, namely 525 records as training data and 175 records as testing data. Based on the results of calculations on the Naïve Bayes method, it is obtained True Positive (TP): 38, True Negative (TN): 113, Positive False

(FP): 15 and False Negative (FN): 9. So that overall accuracy can be obtained based on three parameters can be shown in table 2.

Table 2. System Testing Results

TESTING	ACCURACY	PRECISION	RECALL	F-MEASURE
Comprehensive	86%	72%	81%	76%
Transportation	90%	93%	88%	90%
Hospital	88%	93%	60%	73%
Market	94%	88%	100%	94%
Average	**90%**	**86%**	**82%**	**83%**

Based on the results of system testing carried out both comprehensively (overall parameters) and individually between parameters, results above 50% are obtained, in each parameter of Accuracy, Precision, Recall and F-Measure testing. The average results of system testing were obtained, for Accuracy: 90%, Precision: 86%, Recall: 82% and F-Measure: 83%. All values obtained from the average system testing result above 70%. So, with this shows the system performance has run very well in conducting data classification.

CONCLUSION

In this research has produced a system or application that is able to classify complaints or not complaints based on post data on twitter of three parameters (transportation, hospital and market). Processing of learning machines using the Naïve Bayes method. This system has been able to conduct a series of preprocessing processes as the preparation stage for data input which includes data cleaning, tokenization, POS, stemming, feature extraction and tweet classification. The collection of original data from Twitter is 1000 data, after passing the data cleaning process it is obtained 700 datasets used in the process of machine learning to predict complaints or not complaints on public services.

System testing is done twice, which is overall (comprehensive) and separately (in each parameter). The best test results when done separately. As indicated by the results of transportation accuracy, which is 90%, hospitals are 88% and the market is 94%.

Evaluation of system performance in terms of classification can be seen based on the results of testing of three parameters. So that the average results of the two system tests were obtained for Accuracy: 90%, Precision: 86%, Recall: 82% and F-Measure: 83%.

REFERENCES

Amalia, M. A. Bijaksana, and D. Darmantoro. (2017). “A Framework for Sentiment Analysis Implementation of Indonesian Language Tweet on Twitter A Framework for Sentiment Analysis Implementation of Indonesian Language Tweet on Twitter,” in *International Conference on Computing and Applied Informatics*.

Effendy, A. Novantirani, and M. K. Sabariah. (2016). “Sentiment Analysis on Twitter about the Use of City Public Transportation Using Support Vector Machine Method,” *Intl. J. ICT*, vol. 2, no. 1.

Haddi, X. Liu, and Y. Shi. (2013). “The Role of Text Pre-processing in Sentiment Analysis,” *Procedia Comput. Sci.*, vol. 17.

Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao. (2011). *“Target-dependent Twitter Sentiment Classification,”.*

K. Kim, M. J. Park, and J. J. Rho. (2015). “Effect of the Government ’ s Use of Social Media on the Reliability of the Government : Focus on Twitter,” no. April 2015.

Kurniawan, M. A. Fauzi, and A. W. Widodo.(2017). *"Twitter News Classification Using the Improved Naïve Bayes Method,"* vol. 1, no. 10.

Lee, D. Palsetia, R. Narayanan, M. A. Patwary, A. Agrawal, and A. Choudhary. (2011). *“Twitter*

Trending Topic Classification,”.

Li, Z. Ma, and H. Chen. (2014). “QODM: A query-oriented data modeling approach for NoSQL databases,” in *Proceedings - 2014 IEEE Workshop on Advanced Research and Technology in Industry Applications, WARTIA 2014*.

Weissbock, A. A. A. Esmin, and D. Inkpen. (2013). *“Using External Information for Classifying Tweets.*

Yu. (2002). “Toward an Incremental Democracy and Governance : Chinese Theories and Assessment Criteria,” *New Polit. Sci.*, vol. 24, no. 2.