PROC. INTERNAT. CONF. SCI. ENGIN.
Volume 4, February 2021
Page 171-175

ISSN 1504607797
E-ISSN 1505707533

Application of The Naive Bayes Classifier Method In The Sentiment Analysis of Twitter User About The Capital City Relocation

Syafa'at Adi Nugraha[1], Maria Ulfah Siregar[2]
[12]Informatics Department, [2]Informatics Department Magister Program, Faculty of Science and Technology, UIN Sunan Kalijaga,
Jl. Marsda Adisucipto No 1 Yogyakarta 55281, Indonesia. Tel. +62-274-540971, Fax. +62-274-519739. [1]Email: [1]aatadinugraha@gmail.com,
[2]maria,siregar@uin-suka.ac.id

Abstract: The president's official decision regarding the relocation of the capital city, which previously was in Jakarta and will be moved to East Kalimantan, will be a big issue and generate a lot of debate from both the agree and the disagree Indonesian citizen because this project will require a lot of funds and will have massive impacts on various sectors. This study uses 1290 tweet data consisting of 686 negative sentiments and 603 positive sentiments. These data were used as training data to create evaluation models using confusion matrix and split validation techniques. In this study, it is used TF-IDF word-weighted feature extraction with the Naive Bayes method. The result from the experiments, which was carried out on the best accuracy by splitting data 90:10 for training and testing respectively, is 76.74%. The model that has been made is implemented in 1115 test data resulting in 799 negative sentiments and 316 positive sentiments.

Keywords: capital city, movement, naive bayes, TF-IDF, split validation, confusion matrix.

Abbreviations: TF-IDF

Running Title: Application of The Naive Bayes Classifier Method

INTRODUCTION

The current capital city, Jakarta, is already densely populated, as the center of government, there are various kinds of trade, there are many industries, tourism and chaotic spatial planning and many buildings, and many contradictory land uses occur (Yahya, 2018). The year 2019 was a speculative year because the president gave a discourse on moving the capital city and in the same year it was also officially announced that the former capital city of Jakarta would be moved to East Kalimantan. The official presidential decree that has been announced will be a big issue and could cause a lot of debate regarding moving the capital city because the costs incurred are not small and will affect various sectors such as development projects or budgets for various other projects in other regions. If the capital city is moved, this project needs funds which are clearly not insignificant, it will cause problems related to infrastructure problems and how the type of arrangement will be, including if the movement of the capital is implemented, there will be an extraordinary change of order (Toun, 2018).

The majority of people have used social media, especially twitter, often as a place for people to provide opinions or tweets for criticism and suggestions of something. In this study, we use a sample of tweet data from several Twitter users. Sentiment analysis is used for this research to determine public sentiment regarding the relocation of the capital. Data for this analysis is obtained from tweets from Twitter. We chose to use these data because a tweet has the meaning as an expression or opinion of an individual.

From the research results, information on topics that are often disclosed by the community can be obtained so that the government can respond according to the expectations of the community and can study them appropriately so as to produce good changes for the Indonesian people, especially the areas affected by this project. In addition, the existence of real research examples can be used as information that this capital city relocation program has pros and cons results in the community which are taken from the tweet data of several Twitter users. In this case using several samples of tweets from Twitter users can be used as an illustration of a government program or policy that gets approval or rejection from the public. Negative sentiment indicates that there is resistance and positive sentiment indicates support. The method used is the Naive Bayes Classifier (NBC) which has several advantages including simple, fast and high accuracy (Taheri & Mammadov, 2013). The Naive Bayes Classifier method for classification or categorization of text uses word attributes that appear in a document as the basis for classification. From the data obtained, it will be processed using the Python programming language and will contain information with positive and negative tendencies, and will contain data on words that are often written by several community samples via social media especially twitter.

MATERIALS AND METHODS

Study area

The research method used in this research is the Naïve Bayes Classifer method to classify the analysis of public sentiment on the relocation of the new capital on Twitter social media. The Naïve Bayes algorithm is a form of data classification using probability and statistical methods. In its use, Naive Bayes uses a branch of mathematics known as probability theory to find the greatest probability of possible classifications by looking at the frequency of each classification in the training data.

Procedurs

The Object of Research

The data used in this study used data that was crawled by tweeting from January 2020 to June 2020 with random data collection times with the keyword moving the capital city, the new capital, the capital of Kalimantan. In the data collection process using the Twitter API. The data that can be collected is 5356 tweets. Then the data is divided into 2 parts, namely 1290 training data and 1115 test data for the automatic classification process using the model that has been made.

RESULTS AND DISCUSSION

Several stages that are carried out in this research are as follows:

1. Literature study

The literature study stage is the stage of learning about the theories that will be used in research so that the research process runs well and produces accurate results. The theory related to this research is the theory related to the Naive Bayes Classifier algorithm to analyze the tweet data.

2. Data Collection

Data collection by data crawling, using the Twitter API which previously registered on the Twitter website which is used to obtain large amounts of tweet data. Data taken with the keyword moving the capital city, new capital, and the capital of Kalimantan by getting a number of 5356 tweets.

3. Data Selection and Labeling

The data collection method used is the Simple Random Sampling method. Select data by selecting which data to use and which data is not used, eliminating duplication of data. Then separate the data into test data and training data.

In the process of determining the decision of the labeling results, it is obtained from the largest percentage of the results of three trainers in analysing a sentiment of a tweet. It is depicted in **Table 1**

Table 1. Example of tweet data

Tweet	*Trainer 1*	*Trainer 2*	*Trainer 3*	Result
@geloraco kalau negara gk ada uang gk usa pindah ibu kota.	-1	-1	-1	-1

4. Preprocessing Data

The data obtained is still dirty. Some of the elements are not used, thus cleaning is necessary to perform. It is conducted in the pre-processing stage. Following are the preprocessing stage (Bestari et al., 2019):

a. Cleansing and Casefolding

This process is done by deleting URLs and emails, deleting special twitter characters begins with the process of deleting Twitter special characters such as hashtag (#hashtag), username (@username), and special characters, removing symbols

b. Converting Slangword

Replacing non-standard words (slangwords) to conform to Indonesian standards, this process replaces non-standard words in tweets into standard words that are already known in Indonesian.

c. Stopword Removal

Words that are deemed unimportant or do not describe the contents of the document (stopword list) will be deleted or called the stopword removal process, which means words that are not included in the stopword list will then be converted into their basic form.

d. Stemming

The process of removing affixes such as prefixes, suffixes, and confixes that exist in each word, the affixes will be removed so that each word turns into a basic word.

ANALYSIS AND EVALUATION

Analysis

Data that has gone through the labeling and preprocessing processes will be processed using the Naive Bayes algorithm, then the feature extraction process will be carried out using TF-IDF word weighting. In the calculation, the TF-IDF algorithm calculates the weight of the document with a formula:

Wtd = tftd x IDFt

$$\text{Wtd} = tf(t,d)\ x\ log\ \frac{N_d}{df_t} + 1 \qquad (1)$$

The following is an example of TF-IDF calculations presented in **Table 2**

Table 2. Example of TF-IDF calculation

term	TF		TF(DF)	IDF	IDF + 1	TF-IDF	
	DT1	DT2				DT1	DT2
atas	0	1	1	0,30103	1,30103	0	1,30103
bekas	0	1	1	0,30103	1,30103	0	1,30103
biar	0	1	1	0,30103	1,30103	0	1,30103
cina	1	0	1	0,30103	1,30103	1,30103	0

Furthermore, the application of the Naive Bayes Algorithm (Jurafsky,D & Martin, 2019), namely:

1. Trainning data

a. Calculate the probability of all documents in each class with the following formula:

$$P(c) = \frac{N_c}{N_{document}} \quad (2)$$

Negative class :

$$P(negative) = \frac{1}{2} = 0,5$$

Positive class :

$$P(positive) = \frac{1}{2} = 0,5$$

b. Calculates classification probability

$$P(w_i|c) = count\ \frac{w_i c+1}{\sum w \epsilon\ V\ (count(w,c))+|v} \quad (3)$$

While an example for the calculation of Laplace smoothing is presented in **Table 3**

Table 3. Example of *laplace smoothing* calculation

term	∑W kata t		IDF+1	Laplace smoothing	
	negatif	positif		negatif	positif
atas	0	1,30103	1,30103	0,022508	0,046361
bekas	0	1,30103	1,30103	0,022508	0,046361
biar	0	1,30103	1,30103	0,022508	0,046361
cina	1,30103	0	1,30103	0,051791	0,020148

1. Testing data

After completing the process of looking for the probability of each word for each class, then multiplying all class variables based on sentiment. The maximum value obtained indicates that the document is included in that classification.

In this study, data modeling was carried out using the Naive Bayes method with TF-IDF word-weighted feature extraction with a total of 1290 data with 686 negative labels and 603 positive labels.

Evaluation

In this study using a split validation technique with confusion matrix. Split validation serves as a test of the accuracy of the results of learning training data (Untari, 2010) while the Confusion matrix is information about actual and prediction given by the classifier (Setianingrum et al., 2017). In this study, using 1290 data then carried out with several data splits. The following is the distribution of the split data tested is presented in **Table 4.**

Table 4. Split test data sharing

Split data	The amount of data		Result
	Train data	Test data	
90:10	1160	129	76,74%
80:20	1031	258	75,58%
70:30	902	387	73,13%

Following are the results of calculating accuracy using confusion matrix on a 90:10 split data presented in **Figure 1**

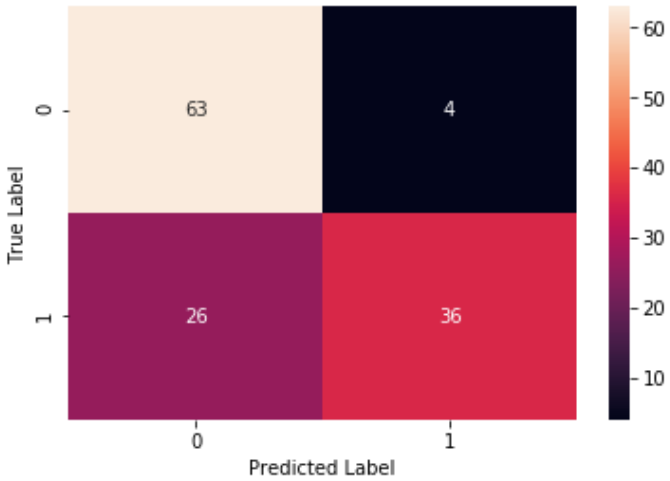

Figure 1. Confusion matrix with data split 90:10

From the results obtained in **Figure 1**. Accuracy can be calculated using the following equation:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (4)$$

$$Accuracy = \frac{36+63}{36+4+26+63}$$

$$= 0{,}7674 \times 100\% = 76{,}74\%$$

Furthermore, the results of the confusion matrix on the 80:20 split data are presented in **Figure 2**

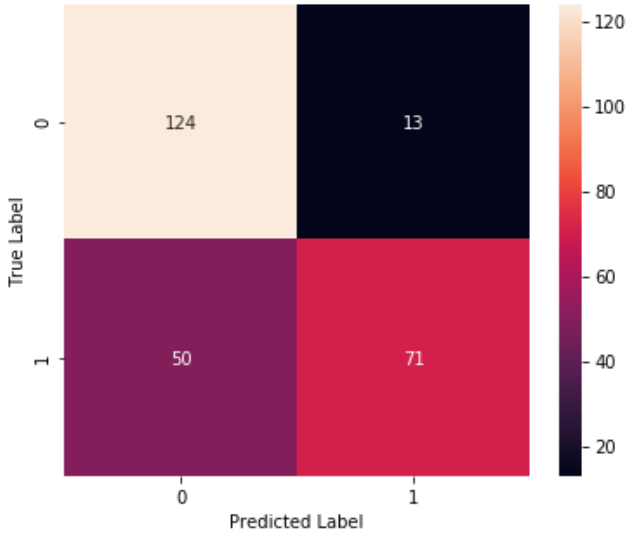

Figure 2. Confusion matrix with data split 80:20

As for the results of the confusion matrix on the 70:30 data split is presented in **Figure 3**

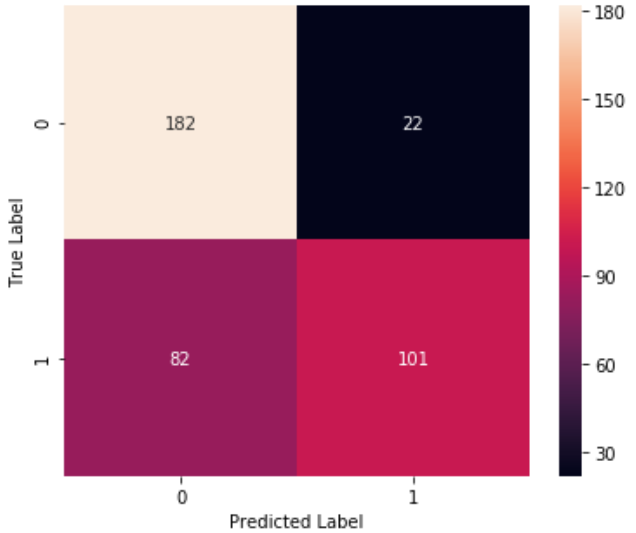

Figure 3. Confusion matrix with data split 70:30

So the conclusion that can be drawn using the split validation technique, namely using split data 90:10 gets the highest accuracy.

IMPLEMENTATION

The classification process uses the Naive Bayes method and weighting of the TF-IDF words on the test data that has been prepared as many as 1290 data. The implementation process uses a model with split data testing 90:10 which has the highest accuracy results with 1115 new data resulting in 799 negative sentiments and 316 positive sentiments resulting in 71.7% negative sentiment and 28.3% positive sentiment . The following is a data visualization of the implementation results is presented in **Figure 4**

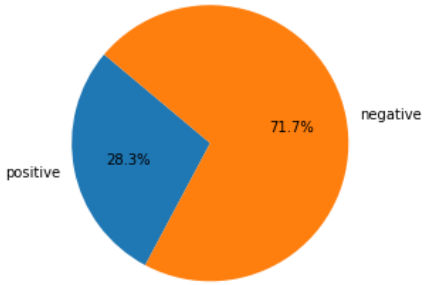

Figure 4. Percentage of test data

Meanwhile, the results of words or topics that often appear in positive sentiment include Kalimantan, Jakarta, wake up, start, and Jokowi. Whereas in negative sentiment, the words that often appear include the state, people, projects, covid, corona, and pandemic. The next visualization, which is based on the sentences that most often appear on negative sentiment, is presented in **Figure 5** while for positive sentiment is presented in **Figure 6**

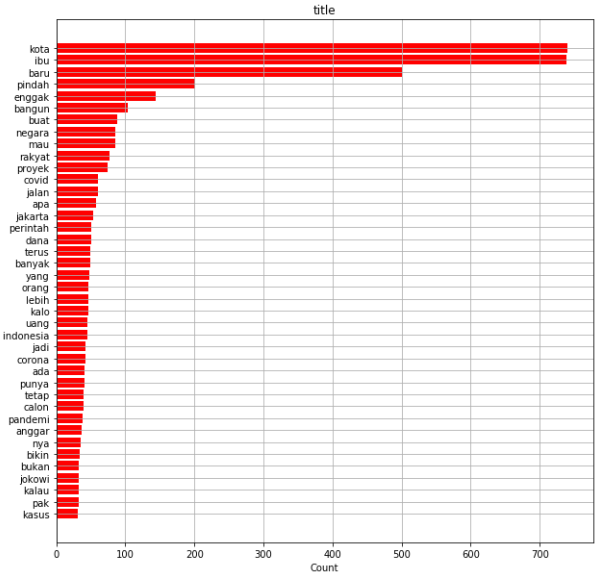

Figure 5. Result of negative word

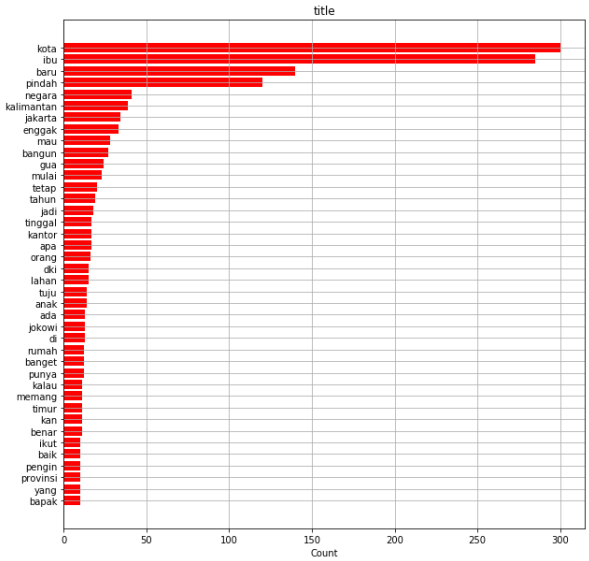

Figure 6. Result of positive word

Discussion

In this study using tweet data from twitter then labeled manually with 3 trainers then analyzed using TF-IDF word-weighting feature extraction then analyzed using the Naive Bayes method, after that evaluation using confusion matrix and split validation. After obtaining the results, the accuracy is then implemented in data testing and the results of the presentation between negative and positive tweets are obtained and the results are in the form of words that often appear or topics that are often written about.

CONCLUSIONS

Based on research conducted from 1290 training data using split validation, it produces split data with the highest result, 90:10, which is calculated using the Naive Bayes method of confusion matrix and TF-IDF word weighting produces 76.74% accuracy. The results of the implementation of 1115 test data resulted in 799 negative sentiments and 316 positive sentiments. The implementation carried out resulted in 71.7% negative sentiment and 28.3% positive sentiment. So the result is that the majority gives negative sentiment or disapproves of the process of moving the new capital city. Topics that often appear on positive sentiment include Kalimantan, Jakarta, wake up, start, and Jokowi. Whereas in negative sentiment, the words that often appear include the state, people, projects, covid, corona, and pandemic.

ACKNOWLEDGEMENTS

Gratitude to all of my family, all of my friend and anyone who have provided encouragement, support and prayer.

REFERENCES

Bestari, D. P. B., Saptono, R., & Anggrainingsih, R. (2019). ACADEMIC ARTICLES CLASSIFICATION USING NAIVE BAYES CLASSIFIER (NBC) METHOD. *ITSMART: Jurnal Teknologi Dan Informasi*, *7*(2), 74–81.

Jurafsky,D & Martin, J. H. (2019). Naive Bayes and Sentiment Classification. In *Speech and Language Processing*.

Setianingrum, A. H., Kalokasari, D. H., & Shofi, I. M. (2017). Implementasi Algoritma Multinomial Naive Bayes Classifier. *JURNAL TEKNIK INFORMATIKA*, *10*(2), 109–118.

Taheri, S., & Mammadov, M. (2013). Learning the naive bayes classifier with optimization models. *International Journal of Applied Mathematics and Computer Science*, *23*(4), 787–795. https://doi.org/10.2478/amcs-2013-0059

Toun, N. R. (2018). Analisis Kesiapan Pemerintah Provinsi Kalimantan Tengah dalam Wacana Pemindahan Ibu Kota Negara Republik Indonesia ke Kota Palangkaraya. *Jurnal Academia Praja*, *1*(01), 129–148.

Untari, D. (2010). Data Mining untuk Menganalisa Prediksi Mahasiswa Berpotensi Non-Aktif Menggunaka Metode Decision Tree C4.5. *Fakultas Ilmu Komputer Universitas Dian Nuswantoro*.

Yahya, H. M. (2018). *Pemindahan Ibu Kota Negara Maju dan Sejahtera*. *14*(01), 21–30. https://doi.org/10.23971/jsam.v14i1.kemerdekaan