Twitter sentiment analysis of the relocation of Indonesia's capital city

Edi Sutoyo, Ahmad Almaarif
Department of Information Systems, Telkom University, Indonesia

Article Info

Article history:

Received Dec 29, 2019
Revised Mar 8, 2020
Accepted Mar 23, 2020

Keywords:

K-nearest neighbor
Logistic regression
Naïve Bayes classifier
Opinion mining
Relocation of the capital city
Sentiment analysis
Support vector machine

ABSTRACT

Indonesia has a capital city which is one of the many big cities in the world called Jakarta. Jakarta's role in the dynamics that occur in Indonesia is very central because it functions as a political and government center, and is a business and economic center that drives the economy. Recently the discourse of the government to relocate the capital city has invited various reactions from the community. Therefore, in this study, sentiment analysis of the relocation of the capital city was carried out. The analysis was performed by doing a classification to describe the public sentiment sourced from twitter data, the data is classified into 2 classes, namely positive and negative sentiments. The algorithms used in this study include Naïve Bayes classifier, logistic regression, support vector machine, and K-nearest neighbor. The results of the performance evaluation algorithm showed that support vector machine outperformed as compared to 3 algorithms with the results of Accuracy, Precision, Recall, and F-measure are 97.72%, 96.01%, 99.18%, and 97.57%, respectively. Sentiment analysis of the discourse of relocation of the capital city is expected to provide an overview to the government of public opinion from the point of view of data coming from social media.

Corresponding Author:

Edi Sutoyo,
Department of Information Systems,
Telkom University,
Bandung, West Java, 40257, Indonesia.
Email: edisutoyo@telkomuniversity.ac.id

1. INTRODUCTION

The capital city in a country plays a very strategic role. This is because the capital city in a country can be multifunctional, that is, as the center of politics and government, the center of business and economic activity, as well as the center of all that, characterizes the overall character of a country. Broadly speaking, it can be concluded that the picture of a country can be seen from how its capital city is. Likewise, Indonesia has a capital city which is one of the many big cities in the world called Jakarta. The role of Jakarta in the dynamics that occur in Indonesia is very central because almost 70% of the amount of money in Indonesia only revolves in Jakarta. This shows that Jakarta, in addition to functioning as a political and government center, is also a business and economic center that drives the economy in Indonesia.

The relocation of the capital city, together with the development of the state and nation, has become an important part of the formation of post-colonial states. There have been national debates and major projects on this issue in many countries in Asia (i.e. Indonesia, India, Malaysia, Sri Lanka, and Pakistan), Africa (i.e. Ivory Coast, Tanzania, Malawi, and Zimbabwe) and South America (i.e. Brazil, Argentina, and Costa Rica). However, over time only a few countries have carried out actual relocation and most projects have been postponed indefinitely [1].

Based on the many countries mentioned above, history has recorded that there are several countries that have succeeded in moving their capital cities with a variety of processes and backgrounds, such as India, Australia, Brazil, Myanmar, Pakistan, Kazakhstan, and Nigeria [2, 3]. Even so, the discourse of moving the capital city generally offers more warning than encouragement. A warning often thrown at the government so far is the cost to realize capital city relocation. For example, when Brazil moved its capital city from Rio de Janeiro to Brazil in 1960. The reason for the relocation is due to the high population density in Rio de Janeiro and the high level of traffic congestion [4]. Reporting from Ultimosegundo [5], the costs spent by the Brazilian government were estimated to exceed USD 1.5 billion at that time, or the equivalent of USD 83 billion if using the exchange rate assumption USD in 2010. Likewise, Myanmar has moved the capital city from Yangon to Naypyidaw in 2005. Reporting from theglobalist [6], the cost spent by the government when led by Than Shwe at that time was estimated to touch 5 billion US dollars. Another example is Kazakhstan, a country whose territory is on two continents, namely Asia and Europe, which cost more than 400 million US dollars to move its capital city from Almaty to Astana in 1998 [4]. The estimated cost of moving the capital city by Kazakhstan is only about 16 percent of the total state revenue at that time was USD 2.47 billion-11 percent of Kazakhstan's total gross domestic product (GDP) of USD 22.13 billion. The reason Kazakhstan moved its capital city to Astana (currently Nur Sultan) is that the location of the previous capital city Almaty was too close to the Chinese border, so it was considered a threat to the political, cultural and economic aspects [7].

The relocation of the capital city is actually not new for Indonesia. Historically, several cities had been the capital city of Indonesia including Yogyakarta, Bukittinggi in West Sumatra, Bireuen District in Aceh [8]. At present, the government is back to discussing the relocation of the capital city because Jakarta is deemed no longer worthy of being a capital city for a country as large as the Republic of Indonesia. In addition, its location further to the western part of Indonesia is blamed for the high level of inequality between regions in the country. Therefore, it is currently being discussed to build an extraordinary megaproject, namely the relocation of the capital from the original in the city of Jakarta to other areas that are considered more potent and have a better regional carrying capacity. In general, countries that have larger land areas tend to have separate political and economic business centers, such as the United States, Australia, Malaysia, Turkey, and etc. [3].

There are several reasons for the government plans to relocate the capital city of the Republic of Indonesia outside of Java Island, one of which is related to the population in Jakarta, which is not decreasing every year but increased significantly [9], it is caused by all the centers of activity in Jakarta such as government centers, economy, business, education, etc. that causes the population of Jakarta to be increasingly crowded. That also causes the availability of clean water in Jakarta to be getting worse [10]. Another reason, according to the study is related to the geographical condition of Jakarta which is in the Ring of Fire, which means it is in a disaster-prone circle [11]. Moreover, the capital city must also be rescued from threats due to the increasingly mismanaged Jakarta City. This mismanagement includes the inability of the government in the past to anticipate the increasing impact of natural threats. Climate change, which is marked by an increase in sea level, is accompanied by an increase in land subsidence due to exploitative urban development of land and water resources. The rapid increase of business center buildings and offices has been followed by massive groundwater extraction. As a result, flooding is increasingly becoming a serious threat to parts of Jakarta City. Common reasons for moving the capital city are socioeconomic considerations, political considerations, and geographical considerations [12]. These three factors are considered Indonesia is still in the analysis for the relocation of its capital city, not only analysis from within the country but also requires an analysis of the experience of other countries in the world that have relocated their capital city. The experience of these countries will be able to provide input and considerations that can be used as more appropriate analytical material to study problems in Indonesia.

However, the discourse of moving the capital city offers more warning than encouragement, so that in addition to several parties agreeing to relocate the capital city, many parties and communities also disagree about the relocation of this capital city. Relocation of the capital city from Jakarta to Penajam Paser Utara and Kutai Kartanegara as an unnecessary policy, considering that Kalimantan is the lungs of the world, and, as if they understood that the capital transfer project would not really pay attention to the environment, they responded to the issue of relocation should be stopped or if it had to be continued, then do not go to Kalimantan. Analysis of the selection of Kutai Kartanegara as a new capital city later turned out to be weak and even did not match the facts. As an expression that there is minimal potential for disaster. In fact, there was a catastrophic flood that resulted in tens of thousands of residents affected and living in trouble. On the other hand, an earthquake which is said to have never greeted Kalimantan, actually occurred shortly after this issue was rolled out. Moreover, the cost of relocation of the capital city is very expensive. According to government estimates, the cost of moving the capital city to Kalimantan Island could reach USD 33 billion [13]. The fund of USD 33 billion or equivalent to IDR 469 trillion is not a small value because it is equivalent to a quarter of total state revenues throughout 2018 which amounted to IDR 1942 trillion. Not to mention, Indonesia's current financial condition is still not a surplus or a deficit. This year, the budget deficit that the government must cover is estimated at USD 296 trillion. Clearly, the discourse of moving the capital city increasingly adds to the budget burden. Therefore, this needs to be further investigated by knowing public opinion about the discourse of relocation of the capital city from Jakarta to Penajam Paser Utara and Kutai Kartanegara.

At present, the public tends to give their opinions and opinions through various media, including social media. Moreover, the past few years have seen a lot of academic interest in the possibility of using social media to gauge public opinion [14]. One of the effective social media to accommodate opinions about the discourse of moving the capital city is by using Twitter, which is fast in reporting the experiences felt by the society as an evaluation material for related parties. Besides Twitter itself is one of the social media that is familiar to be used by the people of Indonesia [15], which certainly will make it easier to collect opinions compared to conducting surveys or distributing questionnaires. Based on research by Semiocast [16], a social media research institute based in Paris, France, the number of Twitter account holders in Indonesia is the fifth largest in the world and is the third most active country to send Twitter messages (tweets) per day. To find out and determine the tendency of Twitter users to post tweets, it is necessary to do Sentiment Analysis [17]. In the context of social media, Sentiment Analysis is how to analyze people who express their opinions on various topics on social media [18]. Sentiment aims to explain public opinion about products, brands, services, politics, or other topics. Companies, governments, and other fields then use these data to make marketing analyzes, product reviews, product feedback, public services, and government policies. Opinions play an important role as product feedback, services, and other topics. Various text mining and sentiment analysis using classification approaches [19-22] such as Naïve Bayes classifier [23-25], logistic regression [26, 27], support vector machine [28, 29], and K-nearest neighbor [30] has been applied for finding the best result and accuracy. Therefore, in this study, sentiment analysis will be conducted to find out public opinion about the discourse of relocation of the capital city from Jakarta to Penajam Paser Utara and Kutai Kartanegara. In this paper, a comparison of popular classifiers is performed to classify public positive and negative opinions about the discourse of relocation of the capital city using the Naïve Bayes classifier, logistic regression, support vector machine, and K-nearest neighbor. The remainder of this paper is explained as follows: Section 2 explains the research methodology used in this study and introduces machine learning approaches for sentiment classification. Section 3 elaborates the dataset used, the experimental results and performance evaluation. Section 4 concludes our study and discusses future work.

2. RESEARCH METHOD

The purpose of this study is to analyze the performance of Naïve Bayes classifier, logistic regression, support vector machine, and K-nearest neighbors in the classification of tweets in determining positive and negative sentiments regarding the discourse of relocation of the capital city. In this study using Twitter as a source to get a dataset. Figure 1 illustrates the sentiment analysis framework proposed in this study.

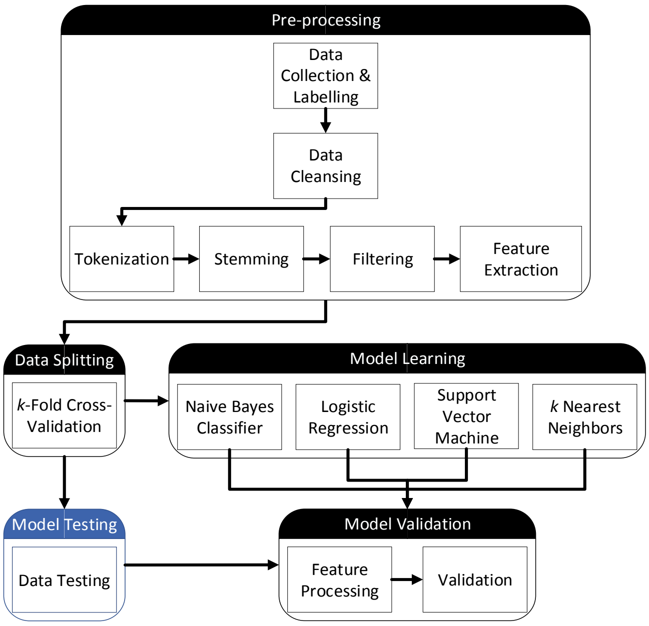


Figure 1. Sentiment analysis framework

The proposed framework consists of 4 (four) main phases: preprocessing, data splitting, learning models, and validation models. The preprocessing phase is related to the process of normalizing the dataset into vector format, which will be input for a classification algorithm that has 6 (six) steps, namely, data collection and labeling, data cleansing, tokenization, stemming, filtering, and feature extraction. The data splitting phase is the process of dividing a dataset into two sets: training data and testing data. Then the learning model stage is continued by conducting learning using each Naïve Bayes classifier algorithm, logistic regression, support vector machine, and K-nearest neighbor so that later performance comparisons can be made. The last step is the validation model: evaluating the performance of each algorithm using 4 (four) evaluation metrics and determining which algorithm has the best performance.

2.1. Sentiment analysis

Sentiment analysis is the process of using text analytics to get various data sources from the internet and various social media platforms and is one of the fields of natural language processing (NLP). Sentiment analysis is the process of using text analytics to get various data sources from the internet and various social media platforms [31, 32]. The goal is to get opinions from users on the platform.

The data can explain public opinion about products, brands, services, politics, or other topics. Companies, governments, and other fields then use these data to make marketing analyzes, product reviews, product feedback, and public services. In order to produce the opinions needed, sentiment analysis must not only be able to recognize opinions from the text. This process, also referred to as opinion mining, also needs to work by recognizing the following three aspects [33]:

Subject : What topic is being discussed
Polarity : Whether the opinion given is positive or negative
Opinion holder : Someone who issues that opinion

2.2. Text processing

Natural language processing (NLP) is a branch of artificial intelligence that focuses on natural language processing. Natural language is a language that is generally used by humans in communicating with each other. The language that is accepted by a computer needs to be processed and understood in advance so that the intent of the user can be understood properly by the computer [34].

Text processing is the process of changing the form of data that has not been structured into structured data in accordance with system requirements. The preprocessing stage is needed to clean the data from noise, homogenize the form of words and reduce the volume of words intended to make the classification method more optimal in the calculation [35]. Preprocessing stages in this study include case folding, tokenization and filtering, stopword removal, stemming [36, 37].

2.3. Case folding

In a document that uses capital letters or the like sometimes, it does not have in common, this can be due to writing errors. In the text preprocessing the case folding process aims to convert all letters in a text document into lowercase letters [37].

2.4. Tokenization

The tokenization stage is the stage of cutting the input string based on each word that makes it up. Tokenization breaks a group of characters in a text into word units, how to distinguish certain characters that can be treated as word separators or not. For example, whitespace characters, such as enter, tabulation, spaces are considered word separators.

2.5. Stemming

Stemming is the process of mapping and breaking down the form of a word into its basic word form. The function of stemming is to eliminate the morphological variations inherent in a word by eliminating the affixes to the word so that later it can get the correct word according to the correct morphological structure [38]. Roughly speaking, the process of changing affixed words into root words by applying language rules to eliminate the affixes. Besides being needed to reduce the number of different document indices, stemming techniques are also for grouping other words that have similar words and meanings/roots but have different forms because they get different affixes. When the basic word of the term has been found so that it can also be found the intensity of the appearance of the term in each document through the indexing process. Indexing is done because a document cannot be recognized directly by an information retrieval system (IRS). Therefore, the document must first be mapped into a representation by using the text inside.

2.6. Filtering

Filtering is the process of selecting important words from the results of tokenization. Filtering is done using a stopword removal algorithm. Stopping or stoplist removal is the process of removing words that do not contribute much to the contents of the document [39]. Stopword removal is used to remove words that often appear and are of a general nature, showing less relevance to the text. Throw out words that often appear but don't have any effect on sentiment extraction. For example, "at", "by", "at", "a", "because" and so forth.

2.7. Term frequency-inverse document frequency (TF-IDF)

Term Frequency-Inverse Document Frequency (TF-IDF) method is a method for calculating the weight of each word that is most commonly used in information retrieval and text mining. The TF-IDF method is a method of giving the weight of a word's terms to documents [40]. This is often used as a weighting factor in information search and text mining. The TF-IDF value increase proportionally based on the number or number of words that appear on the document, but is balanced by the frequency of words in the corpus. Variations of the TF-IDF weighting scheme are often used by search engines as a primary tool in scoring and ranking a document's relevance to a user. This method is one type of term weighting schemes that is popular today because it is efficient, easy and has accurate results. This method calculates the TF and IDF values for each token (word) in each document in the corpus using the following equation:

- Term frequency (TF)

In the case of the term frequency $tf(t,d)$ the simplest way is to use raw frequency in the document, that is, how many times the term t appears in the document d. If raw frequency denoted t as $f(t,d)$, then the simple tf scheme is $tf(t,d) = f(t,d)$. TF can be formulated into:

$$tf(t,d) = 0.5 + 0.5 + \frac{f(t,d)}{max\{f_{t',d:t',d \in d}\}} \tag{1}$$

- Inverse document frequency (IDF)

Inverse Document Frequency (IDF) is a measure of whether the term is common or rare in all documents. This is obtained by dividing the number of documents in the corpus by the number of documents containing the term, and then taking the logarithm of the quotient. The IDF factor of a t-word is given using the following equation:

$$idf(t,D) = log \frac{N}{|\{d \in D: t \in d\}|} \tag{2}$$

The frequency with which words appear in a given document indicates how important that word is in the document. The frequency of documents containing the word indicates how common the word is. Word weight is greater if it appears frequently in a document and smaller if it appears in many documents [41]. Then TF-IDF can be formulated into:

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D) \tag{3}$$

2.8. Naïve Bayes classifier (NBC)

Naïve Bayes is a machine learning method that uses probability calculations rooted in Bayes theorem. Naïve Bayes is based on the Bayes theorem which is used to calculate the probabilities of each class with the assumption that class one is independent (not interdependent). Another definition is Naïve Bayes is a method for predicting future opportunities based on previous experience. Naïve Bayes has a higher level of accuracy and speed when applied to a large value database [42]. The general form of the Bayes theorem is as follow:

$$P(H|X) = \frac{P(H).P(H)}{P(X)} \tag{4}$$

where,
X : Data with unknown classes
H : The data X hypothesis is a specific class
$P(H|X)$: H hypothesis probability based on condition X (posterior probability)
$P(H)$: Hypothesis probability H (prior probability)
$P(X|H)$: Probability of X based on conditions on the hypothesis H
$P(X)$: Probability of X

Naïve Bayes is a simplification of the Bayes method. The Bayes theorem is simplified to:

$$P(H|X) = P(H)P(X) \tag{5}$$

Bayes rules are used to calculate posterior and probability from previous data. The end result will give prior and posterior information to produce probabilities using Bayes.

2.9. Logistic regression

Logistic regression is an algorithm to test the probability that a dependent variable can be predicted with its independent variable. The logistic regression algorithm is used because the analysis with logistic regression does not require the assumption of normality of data on the independent variables. This also means that the logistic regression algorithm is generally used if the assumption of the multivariate normal distribution is not met. The logistic regression can be calculated by using (6),

$$Ln\left(\frac{p}{1-p}\right)B_0 + B_1X \tag{6}$$

where B_0 is a constant, B_1 is the coefficient of each variable. While the value of p can be calculated by (7):

$$p = \frac{e^{(B_0+B_1X)}}{1 + e^{(B_0+B_1X)}} \tag{7}$$

2.10. Support vector machine

Support vector machine (SVM) is a classifier that is now widely used for various classification purposes. In addition to classification, the SVM is also used for regression. The SVM is a binary classifier that divides data into two classes with a hyperplane as depicted in Figure 2. This hyperplane is right in the middle of the two classes with distance d to the nearest data point for each class. d is called the margin, and the data points that are right at the distance d from the hyperplane are called support vectors. The SVM Hyperplane is stated with the following equation:

$$w \cdot x + b = 0 \tag{8}$$

where w is normal from hyperplane, and $\frac{b}{\|w\|}$ is the distance of the hyperplane to the origin.

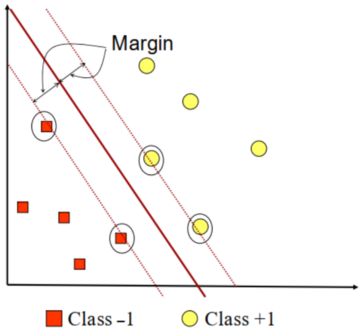


Figure 2. Representation of the SVM method

2.11. K-nearest neighbor

The K-nearest neighbor algorithm (k-NN or KNN) is a data classification method that works relatively in a simpler way compared to other data classification methods. This algorithm tries to classify new data whose class is unknown by selecting the number of data k closest to the new data. The most classes from the nearest data number k are chosen as the predicted class for new data. Similar to clustering techniques [43, 44] on K-Means [45], which is grouping new data based on the distance of the new data to several data/nearest neighbors. k is generally determined in an odd number to avoid the appearance of the same amount of distance in the classification process. First before looking for distance data to neighbors is to determine the value of k. Then, to define the distance between two points namely the point in the training data and the point in the testing data, the Euclidean formula are used. Following is the Euclidean distance used in the KNN algorithm.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{9}$$

KNN is a type of instance-based learning, or lazy learning where this function is only approached locally and all calculations are deferred until classification [46]. The KNN classification method has several stages, the first being the value of k which is the number of closest neighbors which will determine which new query belongs to which class is specified. In the second stage, the nearest k neighbor is searched by calculating the distance of the query point from the training point. The third stage, after knowing the distance of each training point to the query point, then look at the smallest value. The fourth stage takes the smallest value k then look at the class. The most classes are classes from new queries.

2.12. Cross-validation

Cross-validation (CV) is a statistical method that can be used to evaluate the performance of models or algorithms where the data is separated into two subsets namely learning process data and validation/evaluation data [47]. Models or algorithms are trained by training data and validated by testing data. One popular cross-validation method is 10-Fold Cross-Validation. In this technique, the dataset is divided into a number of 10-pieces of random partitions. Then a number of 10-times experiments were carried out, where each experiment used 10-partition data as testing data and utilized the remaining partitions as training data [48].

2.13. Confusion matrix

A confusion matrix is a table used to describe the performance of a classification model on a test data set whose actual labels are known. This allows easy identification of confusion between classes, for example, one class is generally incorrectly labeled as another. The number of true and false predictions is summarized with calculated values and broken down by each class. The confusion matrix shows the ways in which the confused classification determines its class in making predictions. This gives detailed information not only about errors made by the classifier but more importantly the types of mistakes made.

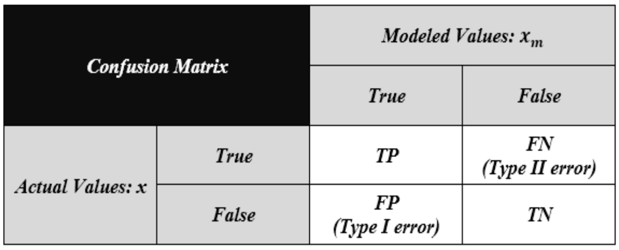
The confusion matrix visualizes the accuracy of the classifier by comparing actual and predicted classes. The binary classifier predicts all data instances of the test dataset as positive or negative. This classification produces four results, namely true positive (TP), false positive (FP), true negative (TN), false negative (FN). TP produces predicted values correctly predicted as actual positives; FP produces predicted values that incorrectly predict true positives. for example, negative values are predicted as positive. Whereas FN produces positive but predicted negative values and TN produces predicted values that are precisely predicted as actual negative [49]. The following Figure 3 explains 4 (four) classes of the confusion matrix classification. From the confusion matrix, measurement metrics can be made to obtain the value of accuracy, precision, recall, and F-measure [50].

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \tag{10}$$

$$Precision = (TP)/(TP+FP) \tag{11}$$

$$Recall = (TP)/(TP+FN) \tag{12}$$

$$F - measure = 2 * \frac{(Recall * Precision)}{(Recall + Precision)} \tag{13}$$



Confusion Matrix		Modeled Values: x_m	
		True	False
Actual Values: x	True	TP	FN (Type II error)
	False	FP (Type I error)	TN

Figure 3. Confusion matrix [51]

3. RESULTS AND DISCUSSION

3.1. Dataset

In this study, the Twitter data collected was user tweets related to the discourse on relocating the capital city. This data was obtained from August 20th, 2019 to September 09th, 2019. The time period was chosen because the discourse on relocation the capital city appeared on August 20th, 2019 and on September 09th, 2019 became the end of data collection because research had entered the stage of writing articles and tweets related to the relocation of the capital city had faded due to being hit by a new trending hashtag that was always changing every day. The process of collecting tweet data is done by utilizing the application interface (API) facility provided by Twitter. The process of collecting data uses several appropriate hashtags and keywords, as shown in Table 1.

The 6 (six) hashtags and keywords (shown in Figure 1) used resulted in 5890 tweets. The six hashtags and keywords were chosen because the hashtags and keywords that produced the most tweets and were the most crowded in the timeline and were also once a trending topic discussing relocation of the capital city. Thus, 6 (six) hashtags and keywords can already represent public opinion about the relocation of the capital city. The sentiment analysis method based on supervised machine learning involves modeling using annotated data and manually labeled. In this study, the process of classifying textual documents (tweets) is divided into two classes/labels, namely positive and negative sentiment classes, which mean tweets, will be given a positive label if they support the relocation of the capital, otherwise, tweets will be labeled negative if they refuse relocation of the capital city.

To reduce computational errors, in the data cleansing stage it is necessary to reduce the number of features by eliminating all textual components that are not useful for classification activities. For this purpose, usernames, full names, user_id, tweet_id, timestamp, various types of characters, punctuation marks, brackets, and sequences in tweets are not useful for classification, intercepted through regular expressions. For example, hashtags (#topic), URLs (http ...), @user (e.g., certain users mentioned in tweets), RT (abbreviations that show retweets), duplicate tweets, tweets from online news are identified and deleted. So, after going through this stage, there are 2891 tweets of ready-to-use datasets, with details of 1638 or equal to 56,66% and 1253 or equal to 43,34% tweets labeled positive and negative sentiments, respectively as illustrated in Figure 4.

Table 1. Hashtags and keywords used in this research

No	Hashtags and Keywords
1	#IbuKotaBaru
2	#pemindahanibukota
3	#PindahIbuKotaUntukSiapa
4	#ibukotapindah
5	Penajam Paser Utara
6	Kutai Kartanegara

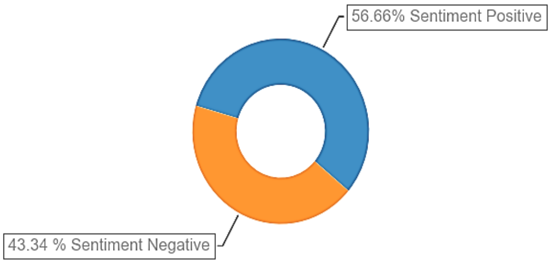


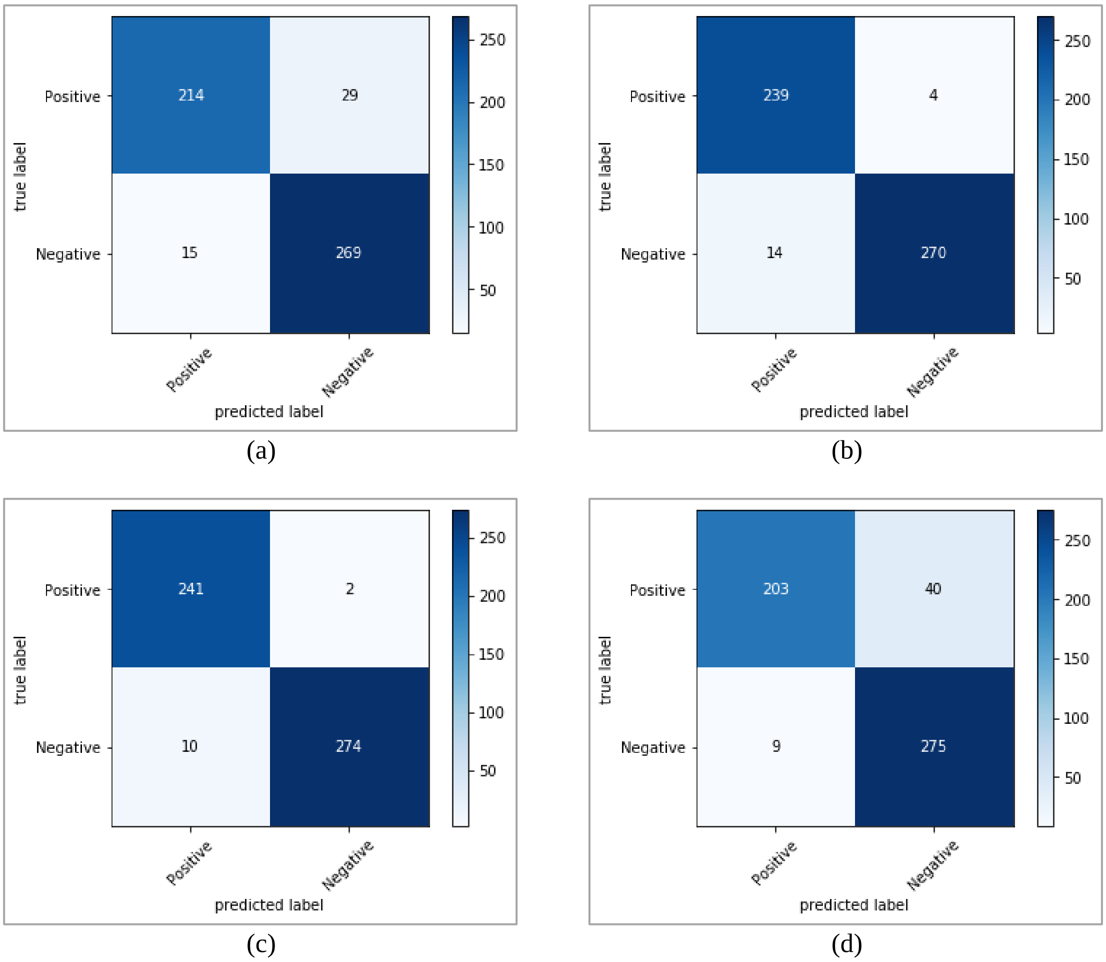
Figure 4. Comparison of the number of tweets that have been labeled

3.2. Experimental results

In this process, 4 (four) algorithms namely Naïve Bayes classifier, logistic regression, support vector machine, K-nearest neighbor are used for the experimental process in order to see the performance of each algorithm. After the experiment process is completed, one more step is needed to determine the quality of the process that has been carried out, namely the evaluation of results. At this stage, the performance of the calculations that have been done will be tested using a confusion matrix.

In the following Figure 5 illustrates the results of the confusion matrix for each algorithm. Figure 5(a) shows the results of the confusion matrix of the Naive Bayes classifier with the results for TP, FN, FP, TN are 214, 29, 15, 269, respectively. Whereas for logistic regression as illustrated in Figure 5(b) obtains results 239, 4, 14, 270. In addition, the results of the confusion matrix for the vector support machine, shon in Figure 5(c), are 241, 2, 10, 274. And the final result for K-nearest neighbor as shown in Figure 5(d) are 203, 40, 9, 275.

From the results of the confusion matrix, the results obtained for 4 (four) evaluation metrics are accuracy, precision, recall, and F-measure. Table 2 below shows the results of 4 (four) evaluation metrics, namely accuracy, precision, recall, and F-measure for each algorithm. The comparison of accuracy performance shows that Naïve Bayes classifier achieves 91.65%, while logistic regression achieves 96.58%, support vector machine achieves an accuracy of 97.72% and K-nearest neighbor achieves an accuracy of 90.70%. Therefore, it can be concluded that support vector machine outperforms the other three algorithms in terms of accuracy. Moreover, support vector machine is superior as compared to the other three algorithms in terms of precision, recall and F-measure with the results of 96.01%, 99.18%, 97.57%, respectively.



(a) (b)

(c) (d)

Figure 5. Confusion matrix results of 4 (four) algorithms,
(a) Naïve Bayes classifier, (b) Logistic regression, (c) Support vector machine, (d) K-nearest neighbors

Table 2. Comparison of the results

Model algorithms	Metrics measurement			
	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
Naïve Bayes classifier	91.65	93.45	88.07	90.58
Logistic regression	96.58	94.47	98.35	96.37
Support vector machine	97.72	96.01	99.18	97.57
K-nearest neighbor	90.70	95.75	83.54	89.23

4. CONCLUSION

In this research, an experiment has been carried out to classify sentiment analysis about government discourse to relocate the capital city. Data sourced from Twitter with the number 2891, with the division of 1638 tweets labeled positive sentiment and 1253 tweets labeled negative sentiment. Four classification algorithms namely Naïve Bayes classifier, logistic regression, support vector machine, K-nearest neighbor are used. The comparison of algorithm performance results has shown that Naïve Bayes classifier, logistic regression, support vector machine, and K-nearest neighbor reached 91.65%, 96.58%, 97.72%, and 90.70%, respectively. This shows that support vector machine outperforms as compared to the other three algorithms in terms of accuracy. Likewise, in terms of precision, recall and F-measure, support vector machine is superior with the results of each of 96.01%, 99.18%, 97.57%, respectively. Sentiment analysis of the discourse of relocation of the capital city is expected to provide an overview to the government of public opinion from the point of view of data coming from social media. Because sentiment analysis can be automated, and therefore decisions can be made based on large amounts of data rather than simple intuitions that are not always true. So that with this research more or less expected to provide some answers about the description of opinions and perspectives of the people about the government's plan to relocate the capital city.

REFERENCES

[1] D. Reva, “Capital city relocation and national security: The cases of Nigeria and Kazakhstan,” University of Pretoria, 2017.

[2] V. Rossman, “Capital cities: Varieties and patterns of development and relocation,” Taylor and Francis, 2016.

[3] P. Ni, M. Kamiya, and R. Ding, “Cities network along the silk road,” Springer, 2017.

[4] E. Illmann, “Reasons for relocating capital cities and their implications,” Charles University, 2015.

[5] Ultimosegundo, “Em dinheiro de hoje, Brasília custaria US$ 83 bilhões,” in English: In today’s money, Brasília would cost US $ 83 billion,” *ultimosegundo.ig.com.br*, 2010. [Online]. Available at: https://ultimosegundo.ig.com.br/brasilia50anos/em-dinheiro-de-hoje-brasilia-custaria-us-83-bilhoes/n1237588758783.html. [Accessed: 12-Dec-2019].

[6] D. Logan, “Why did Myanmar’s generals build a new capital in the middle of nowhere?,” *theglobalist.com*, 2013. [Online]. Available at: https://www.theglobalist.com/myanmars-phantom-capital/. [Accessed: 12-Dec-2019].

[7] M. Arslan, “The significance of shifting capital of Kazakstan from Almaty to Astana: An evalution on the basis of geopolitical and demographic developments,” *Procedia-Soc. Behav. Sci.*, vol. 120, pp. 98-109, 2014.

[8] P. Carey, “Yogyakarta: From sultanate to revolutionary capital of Indonesia. The politics of cultural survival,” *Indones. Circle. Sch. Orient. African Stud. Newsl.*, vol. 14, no. 39, pp. 19-29, 1986.

[9] R. F. Putri, S. Wibirama, Sukamdi, and S. R. Giyarsih, “Population condition analysis of Jakarta land deformation area,” *IOP Conf. Ser. Earth Environ. Sci.*, vol. 148, no. 1, pp. 1-10, 2018.

[10] P. Luo, S. Kang, M. Z. Apip, J. Lyu, S. Aisyah, M. binaya, R. Khrisna, and D. Nover, “Water quality trend assessment in Jakarta: A rapidly growing Asian megacity,” *PLoS One*, vol. 14, no. 7, pp. 1-17, 2019.

[11] F. R. Siegel, “Coastal city flooding,” *Adaptations of Coastal Cities to Global Warming, Sea Level Rise, Climate Change and Endemic Hazards*, Springer, pp. 27-34, 2020.

[12] N. Koch, “The heart of Eurasia? Kazakhstan’s centrally located capital city,” *Centr. Asian Surv.*, vol. 32, no. 2, pp. 134-147, 2013.

[13] K. Lyons, “Why is Indonesia moving its capital city? Everything you need to know,” *www.theguardian.com*, 2019. [Online]. Available at: https://www.theguardian.com/world/2019/aug/27/why-is-indonesia-moving-its-capital-city-everything-you-need-to-know. [Accessed: 10-Dec-2019].

[14] N. Anstead and B. O. Loughlin, “Social media analysis and public opinion : The 2010 UK general election,” *J. Comput. Commun.*, vol. 20, no. 2, pp. 204-220, 2015.

[15] K. M. Carley, M. Malik, M. Kowalchuck, J. Pfeffer, and P. Landwehr, “Twitter usage in Indonesia,” *Center for the Compultational Analysis of Social and Organizational Systems CASOS Technical Report*, pp. 1-54, 2018.

[16] Semiocast, “Geolocation analysis of Twitter accounts and tweets,” *www.semiocast.com*, 2012. [Online]. Available at: https://semiocast.com/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US. [Accessed: 10-Dec-2019].

[17] C. W. Park and D. R. Seo, “Sentiment analysis of Twitter corpus related to artificial intelligence assistants,” *2018 5th Int. Conf. Ind. Eng. Appl. ICIEA 2018*, pp. 495-498, 2018.

[18] A. Pak and P. Paroubek, “Twitter as a corpus for sentiment analysis and opinion mining,” *Proc. 7th Int. Conf. Lang. Resour. Eval. Lr. 2010*, vol. 5, no. 12, pp. 1320-1326, 2010.

[19] I. T. R. Yanto, E. Sutoyo, A. Apriani, and O. Verdiansyah, “Fuzzy soft set for rock igneous clasification,” *2018 International Symposium on Advanced Intelligent Informatics (SAIN)*, pp. 199-203, 2018.

[20] H. Chiroma, S. Abdulkareem, S. A. Muaz, A. I. Abubakar, E. Sutoyo, M. Mungad, Y. Saadi, E. N. Sari, and T. Herawan, “An intelligent modeling of oil consumption,” *Adv. Intell. Syst. Comput.*, vol. 320, pp. 557-568, 2015.

[21] E. Sutoyo and A. Almaarif, “Educational data mining untuk prediksi kelulusan mahasiswa menggunakan algoritme Naïve Bayes classifier,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 1, pp. 95-101, Feb. 2020.

[22] E. Sutoyo, R. R. Saedudin, I. T. R. Yanto, and A. Apriani, “Application of adaptive neuro-fuzzy inference system and chicken swarm optimization for classifying river water quality,” *Proceeding-2017 5th International Conference on Electrical, Electronics and Information Engineering: Smart Innovations for Bridging Future Technologies, ICEEIE 2017*, vol. 2018, pp. 118-122, 2018.

[23] V. Narayanan, I. Arora, and A. Bhatia, “Fast and accurate sentiment classification using an enhanced Naive Bayes model,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8206 , pp. 194-201, 2013.

[24] J. K. Rout, K. K. R. Choo, A. K. Dash, S. Bakshi, S. K. Jena, and K. L. Williams, “A model for sentiment and emotion analysis of unstructured social media text,” *Electron. Commer. Res.*, vol. 18, no. 1, pp. 181-199, 2018.

[25] A. Aninditya, M. A. Hasibuan, and E. Sutoyo, “Text mining approach using TF-IDF and Naive Bayes for classification of exam questions based on cognitive level of bloom’s taxonomy,” *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, pp. 112-117, 2019.

[26] H. Hamdan, P. Bellot, and F. Bechet, “Lsislif: CRF and logistic regression for opinion target extraction and sentiment polarity analysis,” *Proc. of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 753-758, 2015.

[27] K. Bhargava and R. Katarya, “An improved lexicon using logistic regression for sentiment analysis,” *2017 Int. Conf. Comput. Commun. Technol. Smart Nation, IC3TSN 2017*, vol. 2017, pp. 332-337, 2018.

[28] Y. Al Amrani, M. Lazaar, and K. E. El Kadiri, “A novel hybrid classification approach for sentiment analysis of text document,” *Int. J. Electr. Comput. Eng.*, vol. 8, no. 6, pp. 4554-4567, 2018.

[29] C. J. Rameshbhai and J. Paulose, “Opinion mining on newspaper headlines using SVM and NLP,” *International Journal Electrical and Computer Engineering*, vol. 9, no. 3, pp. 2152-2163, 2019.

[30] M. R. Irfan, M. A. Fauzi, T. Tibyani, and N. D. Mentari, "Twitter sentiment analysis on 2013 curriculum using ensemble features and k-nearest neighbor," *International Journal Electrical and Computer Engineering*, vol. 8, no. 6, pp. 5409-5414, 2018.

[31] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundaions Trends® in Inf. Retrieval*, vol. 2, no. 1-2, pp. 1-135, 2008.

[32] G. Vinodhini and R. Chandrasekaran, "Sentiment analysis and opinion mining: A survey," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 2, no. 6, pp. 282-292, 2012.

[33] S.-M. Kim and E. Hovy, "Extracting opinions, opinion holders, and topics expressed in online news media text," *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, no. July, pp. 1-8, 2006.

[34] A. Kao and S. R. Poteet, "Natural language processing and text mining," Springer Science & Business Media, 2007.

[35] S. García, J. Luengo, and F. Herrera, "*Data preprocessing in data mining*," Springer, vol. 72, 2015.

[36] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *1st Int. Conf. Learn. Represent. ICLR 2013 - Work. Track Proc.*, pp. 1-12, 2013.

[37] P. D. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," *J. Artif. Intell. Res.*, vol. 37, pp. 141-188, 2010.

[38] J. B. Lovins, "Development of a stemming algorithm," *Mech. Transl. Compt. Linguist.*, vol. 11, no. 1-2, pp. 22-31, 1996.

[39] R. Baeza-Yates, B. Ribeiro-Neto, and Others, "Modern information retrieval," New York: ACM press, 1999.

[40] A. M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes, "Multinomial naive bayes for text categorization revisited," *Australasian Joint Conference on Artificial Intelligence*, 2004, pp. 488-499.

[41] S. Robertson, "Understanding inverse document frequency: On theoretical arguments for IDF," *J. Doc.*, vol. 60, no. 5, pp. 503-520, 2004.

[42] R. Kohavi, "Scaling up the accuracy of Naive-Bayes classifiers : A decision-tree hybrid," *KDD*, vol. 6319 LNAI, no. 96, pp. 202-207, 1996.

[43] E. Sutoyo, I. T. R. Yanto, R. R. Saedudin, and T. Herawan, "A soft set-based co-occurrence for clustering web user transactions," *TELKOMNIKA Telecommunication Computing, Electronics and Control*, vol. 15, no. 3, pp. 1344-1353, 2017.

[44] E. Sutoyo, I. T. R. Yanto, Y. Saadi, H. Chiroma, S. Hamid, and T. Herawan, "A framework for clustering of web users transaction based on soft set theory," *Proceedings of the International Conference on Data Engineering 2015 (DaEng-2015)*, vol. 520, pp. 307-314, 2019.

[45] N. A. Khairani and E. Sutoyo, "Application of k-means clustering algorithm for determination of fire-prone areas utilizing hotspots in West Kalimantan Province," *Int. J. Adv. Data Inf. Syst.*, vol. 1, no. 1, pp. 9-16, 2020.

[46] S. B. Imandoust and M. Bolandraftar, "Application of k-nearest neighbor (KNN) approach for predicting economic events : Theoretical background," *Int. J. Eng. Res. Appl.*, vol. 3, no. 5, pp. 605-610, 2013.

[47] B. Efron and G. Gong, "A leisurely look at the bootstrap, the jackknife, and cross-validation," *Am. Stat.*, vol. 37, no. 1, pp. 36-48, 1983.

[48] S. Geisser, "The predictive sample reuse method with applications," *J. Am. Stat. Assoc.*, vol. 70, no. 350, pp. 320-328, 1975.

[49] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," *Remote Sens. Environ.*, vol. 62, no. 1, pp. 77-89, Oct. 1997.

[50] C. D. Manning, C. D. Manning, and H. Schütze, "Foundations of statistical natural language processing," MIT Press, 1999.

[51] X. Deng, Q. Liu, Y. Deng, and S. Mahadevan, "An improved method to construct basic probability assignment based on the confusion matrix for classification problem," *Inf. Sci. (Ny).*, vol. 340-341, pp. 250-261, 2016.