Sentiment Analysis Of Indonesian Capital Movement Using The Naive Bayes Algorithm

M Dawa Muzzikri[1)], Susan Dian Purnamasari[2)], Hadi Syaputra[3)]
[1]Fakultas Ilmu Komputer, Universitas Bina Darma,
Jalan jendral A.yani No.3 Palembang Sumatera Selatan, Indonesia
e-mail: dmuzzikri@gmail.com, susandian@binadarma.ac.id, hadisyaputra@binadarma.ac.id

Abstract
Using the hashtags #IKN and #IbuKotaPindah, this study aims to ascertain how Indonesians on Twitter feel about the relocation of the capital. The data is then processed by text processing after being saved in a CSV file. By using the Nave Bayes Classifier method, the text classification process is split into positive and negative sentiment classes. The accuracy of the algorithm is determined by the test results in analyzing the performance of the algorithm using the Confusion Matrix reference and expressing the predictions and actual conditions of the data generated by the algorithm. The f1-score is the average of precision and recall, which are expressed as the positive class at 79 percent and the negative class at 96 percent. Precision is the ratio of the correct amount predicted on the positive class label at 82 percent, negative 95 percent, and recall is the ratio of the correct predicted amount to the overall correct data. Rapid Miner 9.10 is used. Display tab visualizations of the histogram, wordcloud, test data table, and original data table.
Keywords— IKN, Capital Transfer, Naive Bayes Classifier, Kalimantan, Borneo.

1. INTRODUCTION

On Monday 26 August 2019 or more precisely 9 days after Indonesian independence or precisely after 74 years of Indonesia's independence, President-elect Joko Widodo through the official Youtube channel of the Presidential Secretariat announced that the transfer of the new capital city of the Unitary State of the Republic of Indonesia to the Administrative Region of North Penajam Paser Regency and Kutai Kartanegara Regency, East Kalimantan. After the announcement there were many pros and cons that occurred in society [1]. The relocation of the capital will cause an even distribution of development effects [2]. currently the economic level on the island of Java is much higher than other islands in Indonesia [3]. The new capital city can be chosen as a safe area against natural disasters [4], it has the potential to provide large vacant land that can last hundreds of years into the future [5]. The construction of the new capital will increase employment for the surrounding community [3]. On the other hand, a negative opinion was obtained on the relocation of the capital city of Jakarta. The government has to spend around one hundred trillion on infrastructure and transporting central government employees and their families [2]. It is known that the government's current state budget is minimal [4]. In addition, local communities are worried that they will become marginalized [5]. Problems in Jakarta that are the reason, such as floods, garbage, and traffic jams, must be resolved not by relocating the capital [3].

The government's decision to move the capital received a lot of public opinion, including through Twitter. Twitter is the most popular social media for expressing opinions [6]. Expressing opinions on Twitter tends to use informal language and shortened terms [2]. In the past, it was very difficult for the public to express opinions or criticisms because it could only

be done through print media. But now because of the very rapid development of technology, especially communication technology, it is very easy for the public to express opinions, criticisms and suggestions on social media, in this case the community expresses their aspirations for the decision to move the capital city [7].

Sentiment analysis is a method that can solve this problem by changing an unstructured format into a format that has a number of desired classes [6]. Tweets made by the public are a valid source of data for sentiment analysis [8]. There have been many studies using sentiment analysis using the Naive Bayes Classification. Several studies that have used sentiment analysis based on public opinion in connection with government decisions using the Naïve Bayes classification algorithm include sentiment analysis on the implementation of the 2013 curriculum [8] , sentiment analysis on online national exams [9], and sentiment analysis on the application of e-KTP [10]. Referring to previous research, this research is related to government decisions by using the Naive Bayes Classification algorithm. The purpose of this study is to analyze the sentiment of Indonesian-language tweets on Twitter in the form of public opinion on the Indonesian government's decision to move the capital city using the Naïve Bayes classification algorithm. Feature selection using the Term Frequency method [11]. This study analyzes opinions which consist of two classes, namely positive opinions and negative opinions [12]. Data taken from Twitter consists of 974 positive opinion tweets and 433 negative opinion tweets [13]. Rapid Miner software which is quite reliable is used in sentiment analysis [14], which can process the Naïve Bayes classification algorithm; Term Frequency feature selection; and evaluation using Cross Validation and Confusion Matrix [15]. The final result of this study is the average accuracy of the Naive Bayes classification and the selection of Term Frequency features with five comparison ratios between training data and test data, namely the ratios of 90:10, 80:20, 70:30, 60:40, and 50:50. [14].

2. RESEARCH METHODS.

2.1 Data Collection Methods

The following are the techniques used to gather information, analyze it, and resolve issues:

a. Literature Studies
   By gathering publications that are relevant to the author's work, including classification using the Nave Bayes Classifier algorithm, from journals, papers, books, and websites, data is collected using the literature technique.

b. Collection of Tweet Data from Twitter
   Use the Indonesian hashtags # IKN and # IbuKotaPindah on rapid miner tools to get data. The collected data is kept in an excel file with a csv extension.

2.2 Data Analysis Methods

Following is the process for analyzing twitter data to determine its sentiment class :

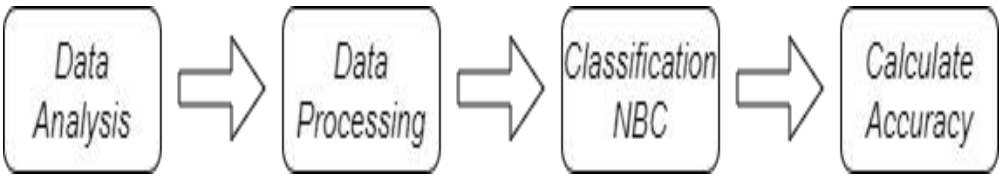

Figure 1. Flow of Sentiment Analysis System

The sentiment flow in Figure 1 can be better understood as follows:

a. Data Source Analysis
The next phase in this study is manually classifying the existing data after using the Twitter API to retrieve information from Twitter, saving it as a Ms.Excel file in csv format, and then import it into a database. There were 1,107 tweets in all, 600 of which were for training and 807 for testing.

b. Data Prepocessing
Tokenizing, transform cases, token filters, and stemming are just a few of the text processing steps that will be used to classify the gathered tweets from Twitter in this procedure. The objective is to arrange the collected data more orderly so that it will be simpler to process.

c. Naïve Bayes Classifier Classification
Each tweet's probability of being included in a favorable or negative opinion will be calculated as part of the classification process using the Naive Bayes Classifier algorithm.

d. Accuracy Results

Calculating the accuracy level is the last step. Since the data under test are already known, determining the system's level of accuracy in the classification process comes after the classification process has been completed. Using a Confusion Matrix to determine accuracy level.

2.3 Naive Bayes Classifier Classification Method

Typically, this process is broken down into many steps. Here are a few examples, among others:

a. Preprocessing text will be applied to unclassified test data. Case folding, stemming, tokenizing, filtering tokens, and stopword erasure are all steps in the preprocessing text stage.

b. The subsequent tweet data will be preprocessed before the frequency of phrases is determined.

c. Then use the formula's equation to determine the Vmap value for each class.

$$Vmap = \underset{Vj \text{ e v}}{argmax} \prod_{i=1}^{n} P(xi|Vj)P(Vj) \quad (1)$$

Information : Vj Tweet categories j=1, 2,...n. Where in this study
j1 = post positive sentiment tweet category,
j2 = negative sentiment tweet categories
P(xi|Vj) Probability xi in categories Vj
P(Vj) Probability of Vj
Based on the tweet's highest Vmap value, the class is determined.

d. Calculate accuracy
using the confusion matrix, determine the system's classification's level of accuracy. Calculations will be made to determine how much of the classification result data is right and how much is incorrect. Use the formula's equation to determine the level of accuracy :

$$\text{Akurasi} = \frac{TP+TN}{TP+FP+TN+FN} \quad (2)$$

Information:

TP : True Positive TN : True Negative

FP : False Positive FN : False Negative

3. RESULT AND DISCUSSION

Tweet information was found on Twitter using the hashtags #IKN and #IbuKotapindah. There are 1,407 tweets in all in Indonesian in the data that was retrieved from Twitter. The information will then be saved in csv format in an excel file. Two different types of data are required for this study: training data and test data. Table 1 shows how training data and test data are separated.

Table 1. Data Sharing

Tweet Sentiment Type	Positive	Negative
Train	205	95
Test	769	338

3.1 Data Reading

The reading of the data into Rapid Miner follows the division of training data and test data. The partitioned training data are in Table 2. The test data have been divided up into Table 3.

Table 2. Train Data

Row	Text	Sentiment
1	mentri “ikn harga mati” — segitu miskin visi	Negative
2	pemerintah kaji kereta gantung jadi transportasi di ikn	Positive
3	demi Bangun ikn	Positive
4	tambahan Pinjaman lgi buat ikn	Positive
5	pemerintah lebih tahu apa yang akan dilakukan. Kereta gantung solusi terbaik ikn	Positive
6	kalau Otaknya belum nyampe	Negative
7	pembebasan Lahan Belum Rampung	Negative
8	dana sudah siap untuk pembangunan ikn	Positive
9	ingat yaa saudara sekalian	Positive
10	Ikn harga mati	Positive

Table 3. Test Data

Row	Text	Sentiment
1	akan seperti apa yaa ikn nanti cant wait	Negative
2	airlangga ajak perusahaan jepang kembangkan kota pintar di ikn	Positive

3	Pembangunan Ibu Kota Negara IKN Nusantara merupakan pembangunan nasional.	Positive
4	ikn joss	Positive
5	padahal dunia sedang alami stagflasi akibat kapitalisme tapi proyek ikn tetap di gas	Negative
6	makanya dulu pak jonan aja menolak loh..makanya beliau gag dipake lg kan..krn beliau pinter dah bs prediksi..ehhh mo diulang loh ga ada kapoknya di proyek mercusuar ikn.hny demi legacy	Negative
7	silahkan jalankan proyek ikn tapi jangan nyusahkan rakyat jugalah lul	Negative
8	subsidi dicabut	Negative
9	kan bisa bangun ikn pake yg 11.000 t itu	Negative
10	pulau kalimantan merupakan lokasi dan jalur transnation crime	Positive

3.2 Text Processing

Text processing is the next step. Tokenization, lowercase letter conversion, punctuation mark removal, number removal, and word deleting (stopword removal) and stemming are the steps conducted at this stage to ensure that the sentiment analysis procedure is unaffected. Figure 2 displays the model for text processing.

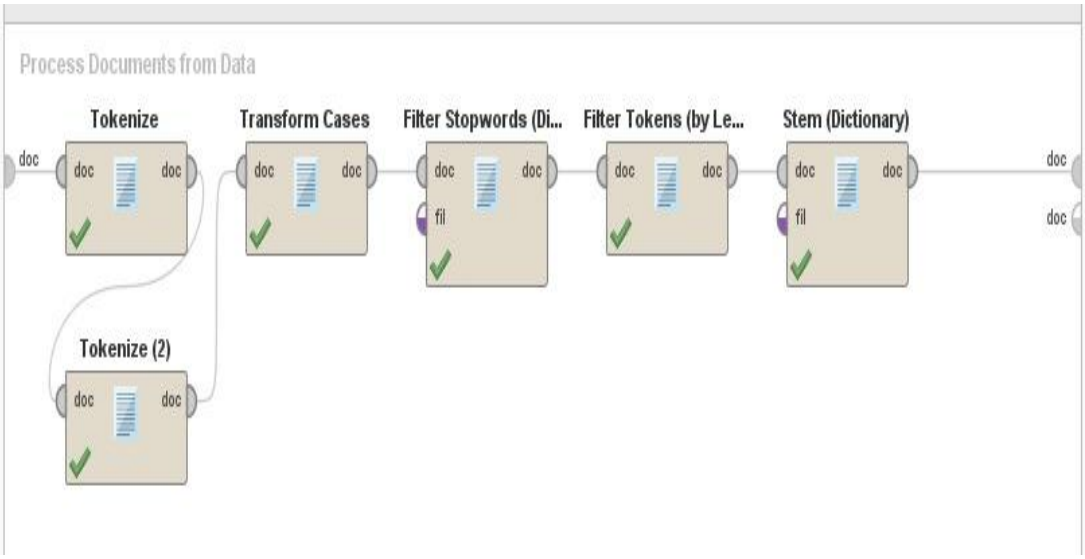

Figure 2. Text Processing Model

3.3 Naive Bayes Model Training

This study used the naive bayes function of the Rapid Miner operators to train the model. Since Naive Bayes evaluates probabilities, a method is needed to provide words that don't present in the sample non-zero probabilities. Figure 3 shows a naive Bayes classifier model, and Figure 4 shows the classification outcomes for test data.

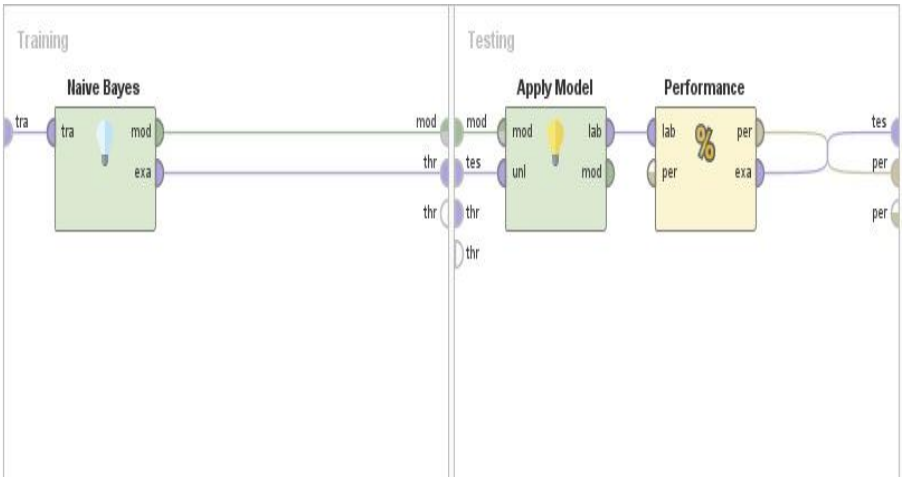

Figure 3. Model Naive Bayes Classifier

accuracy: 87.32% +/- 3.52% (micro average: 87.32%)

	true Negatif	true Positif	class precision
pred. Negatif	747	173	81.20%
pred. Positif	22	596	96.44%
class recall	97.14%	77.50%	

Figure 4. Classification Results

3.4 Calculate Accuracy

The naive bayes classifier method in Figure 5 shows the results of classifying tweets into pro sentiment (positive) as many as 173 documents and neg sentiment (negative) as many as 747 documents based on the testing of 807 tweet data. The accuracy of the naive Bayes classifier method for classifying tweets will be determined from the outcomes of these predictions. to determine the system's classification's level of accuracy using the confusion matrix. The sentiment that was previously known will be compared to the data from the naïve bayes algorithm's categorization. Figure 5 displays the Confusion Matrix's findings.

PerformanceVector

```
PerformanceVector:
accuracy: 87.83% +/- 2.10% (micro average: 87.83%)
ConfusionMatrix:
True:   Negatif Positif
Negatif:        939     202
Positif:        35      772
```

Figure 5. Confusion Matrix Results

According to Figure 6, the results of comparing the sentiment data from the naïve bayes algorithm's classification with the actual sentiment were as follows:

a. True Positive/ TP 772 documents
b. True Negative/TN 939 documents
c. False Positive/FP 35 documents
d. False Negative/FN 202 document

accuracy: 87.83% +/- 2.10% (micro average: 87.83%)

	true Negatif	true Positif	class precision
pred. Negatif	939	202	82.30%
pred. Positif	35	772	95.66%
class recall	96.41%	79.26%	

Figure 6. Naive Bayes Classification Accuracy

3.5 Visualization

The information about the terms that are used most frequently is shown at this stage. Wordclouds and bar charts are the visuals employed.

Row No.	word	in documents	total
1	pembangunan	319	329
2	nusantara	230	233
3	mati	127	142
4	harga	110	121
5	kota	103	104
6	negara	94	95
7	dukung	88	89
8	indonesia	77	78
9	lokal	75	77
10	rakyat	66	74
11	pemerintah	71	71
12	kemajuan	63	64
13	penuh	61	62
14	penduduk	55	57
15	bangun	53	55

Figure 7. Visualization Data

Figure 8. Visualization Wordcloud

The words that appear the most frequently in the tweet data are listed in Figure 7. Figure 8 shows a wordcloud with a list of the words that appear the most frequently. Development, archipelago, dead, price, city, country, support, Indonesia, locals, people, government, progress, full, population, wake up, support, community, head, ensure, infrastructure, authority, pioneer, project, transfortation, jokowi, economy, and president are among the 27 words that appear the most frequently.

3.6 Research Result

The confusion matrix's findings revealed an accuracy of up to :

$$\frac{TP+TN}{TP+FP+TN+FN} = \frac{772+939}{772+939+202+35} = \frac{1.711}{1.948}$$

The naive bayes classifier system correctly classified tweets into positive and negative class attitudes with an accuracy of 87.83, or 87 percent. The precision results for each system class are determined using the confusion matrix results as follows:

a. Positive Class Precission

$$\frac{TP}{TP+FN} = \frac{202}{202+35} = 202 = 82{,}30$$

Positive sentiment accuracy is 82.30 %, or 82.0%, meaning that the rate of class misclassification is 18%.

b. Negative Grade Precision

$$\frac{TN}{TN+FN} = \frac{939}{772} = 939 = 95{,}66$$

Since the sentiment's accuracy result is 95.66 % positive, the class misclassification rate is 5 %.

The accuracy rate of the correct number of positive data classified as positive data and the correct amount of negative data classified as negative data has an impact on the error rate of classifying sentiment classes, according to the results of the positive class precision and the negative class precision. The imbalance of data between the positive sentiment class and the negative sentiment class in the training data results in at least one reference to positive words, which will affect the classification stage for the test data and cause several prediction errors, which in turn influences the error of classification prediction results contained in the confusion matrix table.

4. CONCLUSION

The accuracy of this study's classification of negative and positive sentiment tweets using the Nave Bayes Classifier Method with 600 training data and 804 test data was 78%. Sentiment analysis has shown to be effective in identifying public sentiment about the move of Indonesia's capital to East Kalimantan, particularly among Twitter users. This has aided regular people in learning what other people think of the decision. All users have access to the Rapid Miner application, which is used in this study.

5. SUGGESTION

For further researchers in terms of making analysis of public sentiment on social media, it is hoped that they can apply data mining methods using different algorithms, so that it can be seen whether there are differences in the results obtained when applying other algorithms or can do combinations with different methods or approaches. others to get better research results.

THANK YOU NOTE

The author would like to thank my supervisor, Mrs. Susan Dian Purnamasari, M.kom. who has directed the author to complete this research, as well as thanks to all my friends from the 2018 batch of Bina Darma University Palembang.

REFERENCES

[1] Safra, Icha Adellia, and Eri Zuliarso. "Analisa sentimen persepsi masyarakat terhadap pemindahan ibukota baru di kalimantan timur pada media sosial twitter." (2020).
[2] Taufiq, Muhammad. "Pemindahan ibu kota dan potensi konektivitas pemerataan ekonomi. "Prosiding Seminar Nasional Pemindahan ibu kota Negara. 2017.
[3] Hutasoit, Wesley Liano. "Analisa pemindahan ibukota negara." DEDIKASI: Jurnal Ilmiah Sosial, Hukum, Budaya 39.2 (2019): 108-128.
[4] Septiana, Dwiani, and Sumarlam Sumarlam. "Palangka Raya the Capital City of Indonesia: Critical Discourse Analysis on News about Moving the Capital City from Jakarta." International Seminar on Recent Language, Literature, and Local Cultural Studies (BASA 2018). Atlantis Press, 2018.
[5] Yahya, Muhammad. "Pemindahan ibu kota negara maju dan sejahtera." Jurnal Studi Agama dan Masyarakat 14.1 (2018): 21-30.
[6] Tyagi, Priyanka, and R. C. Tripathi. "A review towards the sentiment analysis techniques for the analysis of twitter data." Proceedings of 2nd international conference on advanced computing and software engineering (ICACSE). 2019.
[7] Castillo, Mendoza, and Marcelo Mendoza. "Poblete,“." Information Credibility on Twitter”, Information Credibility, WWW (2011).

[8] Pamungkas, Dyarsa Singgih, Noor Ageng Setiyanto, and Erlin Dolphina. "Analisis sentiment pada sosial media twitter menggunakan naive bayes classifier terhadap kata kunci â€ œkurikulum 2013â€." Techno. Com 14.4 (2015): 299-314.
[9] Priyono, F., Kanti, S., Dzulfiqar, I., Amirulloh, I., Alvi, A., & Rosiyadi, D., 2016, Analisis Sentimen Media Sosial Opini Ujian Nasional Berbasis Komputer menggunakan Metoda Naive Bayes. Journal of Electrical And Electronics Engineering, No.2, Vol.1.
[10] Mihuandayani, M., Feriyanto, E., Syarham, S., & Kusrini, K., 2018, Opinion Mining pada Komentas Twitter E-KTP Menggunakan Naïve Bayes Classier. Semnasteknomedia Online, No.1, Vol.6, 1-2.
[11] Rofiqoh, U., Perdana, R. S., & Fauzi, M. A., 2017, Analisis Sentimen Tingkat Kepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonesia Pada Twitter Dengan Metode Support Vector Machine dan Lexicon Based Features. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, No.12, Vol.1, 1725-1732.
[12] Antinasari, P., Perdana, R. S., & Fauzi, M. A., 2017, Analisis Sentimen Tentang Opini Film Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naive Bayes Dengan Perbaikan Kata Tidak Baku. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, No.12, Vol.1, 1733-1741.
[13] Fanissa, S., Fauzi, M. A., & Adinugroho, S., 2018, Analisis Sentimen Pariwisata di Kota Malang Menggunakan Metode Naive Bayes dan Seleksi Fitur Query Expansion Ranking. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, No.8, Vol.2, 2766-2770.
[14] Perdana, R.S., 2018, Penerapan Sentimen Analisis Acara Televisi Pada Twitter Menggunakan Support Vector Machine dan Algoritma Genetika sebagai Metode Seleksi Fitur. Jurnal Pengembangan Teknologi Informasi dan Komputer, No.3, Vol.2, 998-1007.