Jurnal Mantik, 7 (2) (2023), ISSN 2685-4236 (Online)

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik

Comparison between naive bayes method and support vector machine in sentiment analysis of the relocation of the Indonesian capital

Imam Rasyidin Muqsith Rizqi Prasetyo[1], Aziz Musthafa[2], Taufiqurrahman[3]

[1,2,3] Informatics Engineering, Science and Technology, University of Darussalam Gontor, Ponorogo, Indonesia

A R T I C L E I N F O

Article history:

Received Apr 7, 2023
Revised Apr 20, 2023
Accepted May 30, 2023

Keywords:

Capital Relocation
Naïve Bayes
Sentiment Analysis
Support Vector Machine

ABSTRACT

Moving the capital city of Indonesia has drawn pros and cons among the public. Therefore, it is important to analyze public sentiment towards moving the Indonesian capital to Kalimantan. In this study, we used data from Twitter and YouTube comments as many as 3895 and 1884 data, starting from 18 May to 6 July 2022. The purpose of this study was to classify public sentiment towards the move of the Indonesian capital city into positive, negative and neutral, as well as compare the results of sentiment analysis using the Naïve Bayes and Support Vector Machine methods. The K-Fold Validation method is used to measure the accuracy of sentiment analysis results. The results of the analysis show that SVM has better accuracy than Naïve Bayes with an accuracy percentage of 0.897 and 0.802 respectively. The resulting comment labels indicated that 56% were positive, 32% neutral, and 11% negative. In this study, we also compared the results of previous studies using the same method, namely Naïve Bayes and SVM. This research can assist the government in evaluating public opinion on the relocation of the Indonesian capital and can be a reference for future researchers in analyzing public sentiment in the future.

Corresponding Author:

Imam Rasyidin Muqsith Rizqi Prasetyo,
Informatic Engineering, Science and technology
University Of Darussalam Gontor,
Jl. Raya Siman, Dusun I, Demangan, Kec. Siman, Kabupaten Ponorogo, Jawa Timur, 63257, Indonesia
Email: imamrasyidinmuqsithrizqiprasetyo@mhs.unida.gontor.ac.id

1. INTRODUCTION

The move of Indonesia's capital from Jakarta to Kalimantan invited mixed responses among the public. Some welcomed it positively, while others viewed it as controversial and contextual. Therefore, it is very important to analyze public sentiment towards the movement of the Indonesian capital. In the digital era, social media is the most widely used platform for people to share their opinions. Therefore, data from social media can be used as an important source of data in analyzing people's sentiments.(Nur Jamal Shaid, 2022)

Several previous studies have been conducted to analyze public sentiment towards the movement of the Indonesian capital using certain classification methods.

Primandani Arsi conducted research using the Naïve Bayes method in 2021. This study resulted in an accuracy of 94.33% in classifying people's sentiments.(Arsi and Waluyo, 2021) in 2021, Primandani Arsi and Retno Waluyo conducted research using the Support Vector Machine method and succeeded in producing an accuracy of 96.68%.(Arsi, Kusuma and Nurhakim, 2021) In the same year, Nabilla Surya Wardani, Alan Prahutama, and Puspita Kartika Sari conducted research using the Naïve Bayes method for Bernoulli and Multinomial models. This study resulted in an accuracy of 93.45%.(Wardani, Prahutama and Kartikasari, 2020) In 2021, Erica Mas'udah, Eka Dyar W., and Amalia Anjani conducted a study using the Naïve Bayes method for each model on Twitter data and managed to produce the best accuracy, namely multinominal Naïve Bayes with an accuracy of 64.40%.(Mas'udah, Wahyuni and Arifiyanti, 2020) In the same year, Tezza Fazzar Tri Hidayat conducted research using the Support Vector Machine method on Twitter data and managed to produce an accuracy of 78.33%.(Hidayat, Garno and Ridha, 2021).

Looking at some of the previous studies, it can be seen that there is a lot of research on sentiment analysis that uses the naive Bayes classifier and support vector machine methods, so the authors conclude that these two methods are methods that are often used for sentiment classification. However, in previous studies there was no discussion about which method is more efficient between the two methods, the authors try to compare the two methods in order to find out which method is more efficient in sentiment classification.

This research has several differences with previous studies by comparing the two methods, namely the Naive Bayes classifier and the support vector machine. This comparison can be seen from the evaluation results using the confusion matrix and validation with k-fold cross validation on the training data obtained from opinions and Twitter comments on YouTube videos by dividing into three classes, namely positive, negative and neutral. It is hoped that this research can contribute to the evaluation of public opinion and become a reference for other researchers in analyzing public sentiment in the future.

This study used data from Twitter and YouTube comments with a total of 3895 data and 1884 data, starting from May 18 to July 6, 2022. This study used the Naïve Bayes method and Support Vector Machine to classify people's sentiments. This study also uses the K-Fold Validation method to measure the accuracy of sentiment analysis results. This research focuses on sentiment analysis of capital city moves with the keywords New Capital City and Capital Relocation.

2. RESEARCH METHOD

2.1. Data Collection

The data used in this study was taken from twitter and youtube comments related to the movement of the Indonesian capital city. Data was taken from May 18 to July 6, 2022. The data taken amounted to 3895 data from twitter and 1884 data from youtube comments.(CNN Indonesia, 2022) This process is done by crawling data using the services of a website called Netlytic which provides features to crawl data. The stages of data collection used include: (a) Determine the data source: At this stage, the data source to be used for sentiment analysis is determined. Data sources should be chosen carefully and should fit the purpose of sentiment analysis. (b) Data collection: At this stage, data or text is collected from a predetermined data source. Data can be retrieved using available tools or APIs, or done manually by collecting text data from these sources. (c) Determine the scope of data: After the data is collected, it is necessary to determine the scope of data to be used in sentiment analysis. The scope of data can include specific times, sources, or topics.

2.2. Data Labelling

In this study, data labeling was carried out manually on 5779 data with three classes divided into positive, negative, and neutral. The labeling process aims to determine the class of tweets related to information or opinions regarding the relocation of the capital, complaints or criticisms, and irrelevant tweets. Once data is manually labeled, it must go through an expert validation process. In this study, the data was validated by Mastufa Siyus Setyowati, a teacher at Muhammadiyah Boarding School Yogyakarta.

Table 1 Data Table

No	Tweets	Label
s	Memenuhi kebutuh rakyt aj gk bs mau bangun IKN. Yg kamu bayangkn cm keuntungn bg kamu.	Negati f
2.	Satu Kata LUAR BIASA. Maju terus Negri INDONESIA. 😍 👨‍🎓 👩‍🎓	Positif
3.	Ngeri desainnya	Netral

2.3. Pre-Processing Data

The preprocessing process is carried out to clean the data from unimportant information and obtain more relevant information. The preprocessing process carried out in this study includes the removal of punctuation, unimportant words, conversion of text to lowercase, removal of word affixes and removal of stopwords.(Yulita, 2021).

Some stages in data preprocessing in sentiment analysis include:(Laurensz and Eko Sediyono, 2021). (a) Cleaning: The first stage is to clean the data of unnecessary or irrelevant characters or signs, such as punctuation marks, emoticons, symbols, or other non-alphanumeric characters. (b) Case Folding: This stage is done to ensure that words in the text are treated equally, both in case (capital or lowercase), word development (stemming), and the removal of unimportant words (stopwords). Word normalization is carried out to ensure that the same words with minor variations are not counted separately in the analysis. (c) Stopword Removal: Stopwords are words that are not important or have no special meaning in the analysis, such as "dan", "atau", "yang", and so on. Eliminating stopwords can help in improving the accuracy of sentiment analysis as it can reduce "noise" or useless information in the data. (d) Tokeninzing: This stage is done to break down text into individual tokens or words, which will later be used as features in sentiment analysis. (e)Stemming: This stage is a technique to reduce the words in a text to a basic form or base word. The goal is to eliminate variations in word forms such as plurals, verbs of different forms (e.g., "berlari" and "berlari"), and variations of words derived from the same root word (e.g., "lari," "berlari," and "lari-lari"). By doing stemming, sentiment analysis can more accurately recognize important words in the text that have the potential to affect existing sentiment.

2.4. Feature Extraction

Feature extraction is performed to obtain important information from the text. In this study, feature extraction was carried out using the TF-IDF method. TF-IDF (Term Frequency-Inverse Document Frequency) is a weighting method often used in text processing and document analysis. TF-IDF is used to judge the importance of a word in a document or corpus of text based on how often it appears in the document and how common the word is in the entire corpus of text.(Atika, Styawati and Ari Aldino, 2022)

Basically, the TF-IDF method calculates two values for each word in the document: Term Frequency (TF) and Inverse Document Frequency (IDF). TF measures how often the word appears in a document, while the IDF measures how common the word is in the entire corpus of text.(Wibowo, 2016).

To calculate the weight of TF-IDF, it first calculates the frequency of occurrence of words in the document (TF), that is, the number of occurrences of a particular word

divided by the total number of words in the document. Then, the IDF value is calculated by dividing the number of documents in the corpus by the number of documents containing the word. Finally, the weight of the TF-IDF is calculated by multiplying the TF value by the IDF value.

With the TF-IDF method, words that appear more frequently in a document will have a higher weight, but words that appear in the entire document with a high frequency will have a lower weight. For example, words like "and", "or", and "to" appear with high frequency in many documents and therefore carry a low IDF weight.(Anggraini, Harahap and Kurniawan, 2021).

The TF-IDF method can be used for various purposes, such as document classification, topic determination, and information retrieval. For example, in document classification, words that have a high TF-IDF weight in a particular category can be used as a feature to distinguish documents from that category from documents from other categories.

2.5. Data Modelling

In this study, the data classification process was carried out using the Naïve Bayes classifier (NBC) method. This method can classify data with simple probability based on Bayes' theorem with high independent characters (Apriyani, 2020). The Naïve Bayes method is widely used in classification techniques, especially in sentiment analysis, such as classification on Twitter using several methods such as Gaussian Naïve Bayes, Multinomial Naïve Bayes, and Bernoulli. The main feature of the Naïve Bayes classification is to derive a strong hypothesis of any condition or event. The calculation of probability categories in Naïve Bayes uses the Bayes algorithm approach with certain equations. This method is suitable for quickly classifying a lot of data and has high accuracy.(Sholihin *et al.*, 2019)

$$P(S|D) = (P(D|S) * P(S)) / P(D)$$

Information:

P(S| D) : probability of sentiment class S on document D
P(D|S) : probability of document D appearing in sentiment class S
P(S) : probability of occurrence of sentiment class S
P(D) : probability of occurrence of document D in all sentiment classes

Explanation: This formula is used to calculate the probability of a document falling into a specific sentiment class. This probability is calculated based on the probability of the document appearing in the sentiment class and the probability of the occurrence of the sentiment class itself.(Lorena., 2016).

In addition to Naïve Bayes, the SVM (Support Vector Machine) method is also one of the classification methods that are often used. SVM uses a linear approach to divide data into classes, which is done by finding the best dividing line between two classes. The SVM method has advantages in classifying data with complex features and has good generalization capabilities.(Monika Parapat and Tanzil Furqon, 2018) The SVM method is also able to handle data with high dimensions and has high accuracy in data classification. In addition, SVM methods can be used in different types of data classification, such as text, image, and sound classification. In its use, SVM requires sufficient training data to obtain an optimal classification model.(Fitriyah, Warsito and Maruddani, 2020)

$$f(x) = sign(w*x + b)$$

where f(x) is a hypothetical function that determines which class to choose, w is the weight vector specified during the training process, x is the input feature vector, and b is biased.(Nanda *et al.*, 2022).

2.6. Model Validation and Evaluation

Data validation is carried out to ensure the reliability of the sentiment classification model that has been built. Validation is performed using the K-Fold Cross Validation method.

K-Fold Cross Validation is a model performance evaluation technique used to avoid overfitting and estimate model accuracy more accurately. This technique takes all available data and divides it into K equal parts. Furthermore, as many as K experiments are carried out, where in each experiment, one of the K parts will be used as test data (testing set) and the other K-1 part will be used as training data (training set). Then, the model will be trained on a training set and tested on a test set to get accuracy from the model.(Nasution and Hayaty, 2019).

The advantage of K-Fold Cross Validation is that the resulting model is more accurate because all data is used for training and testing, so there is no tendency for the model to overfit or underfitting. In addition, this technique also ensures that the model is able to generalize new data, so it is better used on unknown data.

An example of its application to sentiment analysis is to use K-Fold Cross Validation to test the performance of SVM models on tweet data that has been labeled positive and negative sentiment. The tweet data will be divided into K parts, then conducted experiments K times, with each experiment choosing one part as the testing set and the other part as the training set. After completing the experiment K times, the average accuracy of the model will be taken to get a more accurate accuracy value. Thus, the use of K-Fold Cross Validation can help improve the performance of SVM models in sentiment analysis.

3. RESULTS AND DISCUSSIONS

The data used in this study were taken from Twitter and YouTube comments related to the transfer of the Indonesian capital. The data was taken from May 18 to July 6 2022. The data taken amounted to 3895 data from Twitter and 1884 data from YouTube comments. This process is carried out by crawling data using a website called Netlytic which provides features for crawling data.

In this study, manual data labeling was carried out for 5779 data with three classes which were divided into positive, negative, and neutral. The labeling process aims to determine the class of tweets related to information or opinions regarding the relocation of the capital city, complaints or criticism, as well as irrelevant tweets. After the data is labeled manually, the data must go through an expert validation process. In this study, the data was validated by Masfufa Siyus Setyowati, a teacher at the Yogyakarta Muhammadiyah Boarding School.

In this Labeling phase, the results of manual labeling are obtained with positive, neutral, and negative parameters with the number shown in table 2.

Table 2 Labeling Results

No.	Label	Sum
1.	Positive	3269
2.	Negative	1860
3.	Neutral	650

The labeling results shown in table 3 show the results of unbalanced sentiment classes. because of this gap the accuracy results obtained will be low. To overcome this, a balancing step is carried out using the random oversampling method so that the minority data will be duplicated as much as the majority data to produce balanced data. Figure 1 and Figure 2 shows the results of balancing data.

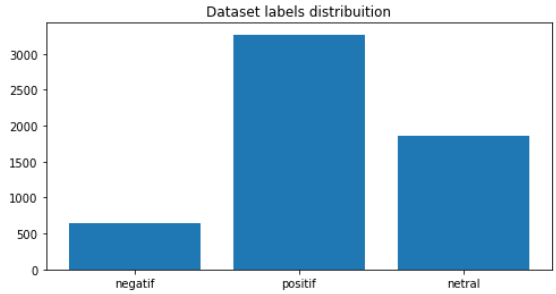

Figure 1 Label Chart Befor Balancing

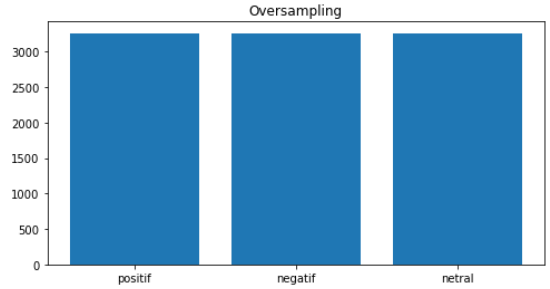

Figure 2 Label Chart After Balancing

Sampling can be seen in Figure 2. The data becomes balanced after making random duplicates of the data with the minority class as much as the difference from the majority data.

After the data is balanced, the next step is the preprocessing stage, this stage is carried out to clean the data from unimportant information and obtain more relevant information. The preprocessing processes carried out in this study include Cleaning, Case folding, Stemming, Stop word Removal, and Tokenizing. Pre-processing results are displayed in the table

Table 3 Preprocessing Results

No.	Tweets	Label
1.	['penuh', 'butuh', 'rakyt', 'aj', 'gk', 'bs', 'bangun', 'ikn', 'yg', 'bayangkn', 'cm', 'keuntungn', 'bg']	Negatif
2.	['ngeri', 'desain']	Netral
3.	['maju', 'negri', 'indonesia', '', '']	Positif

After the data preprocessing process is carried out, then the classification process is carried out using the Naïve Bayes method and Support Vector Machine with evaluation using the Confussion Matrix. The accuracy results for the Naïve Bayes method have a value of 0.802, while the Support Vector Machine method has a value of 0.897. After evaluation, the model will be validated using k-fold cross validation by dividing the data into several parts with the same amount ratio but having a different sample for each part. This model conducts training with training data and testing using testing data as many as the number of folds used. In this research the authors used 10-fod cross validation and obtained the results that will be displayed in table 4.

Table 4 K-fold Details

K-fold	Naïve Bayes	Support Vector Machine
1	0.73	0.75
2	0.72	0.76
3	0.71	0.73

4	0.72	0.77
5	0.75	0.75
6	0.78	0.76
7	0.76	0.72
8	0.74	0.77
9	0.76	0.74
10	0.74	0.76
AVG	0.74	0.75

In the final stage, an analysis of sentiment labels consisting of positive, negative, and neutral is carried out. The analysis showed that there were 56% positive comments, 32% neutral comments, and 11% negative comments. From the results of the study, it can be seen that the Support Vector Machine method produces higher accuracy compared to the Naïve Bayes method. This could be due to Support Vector Machine's better ability to classify complex data. In addition, SVM is also able to handle data that has high dimensions and data that is not normally distributed. In previous research, Hilda Apriani also found that the Support Vector Machine has better accuracy than the Naive Bayes classifier. (Apriyani, 2020).

Meanwhile, the Naïve Bayes method has the advantage of faster and more efficient performance, and requires fewer computing resources. This advantage is very useful in processing large and complex data. In the sentiment analysis of the relocation of the Indonesian capital, there was a majority of positive comments at 56%, followed by neutral comments at 32% and negative comments at 11%. This shows that the majority of people have a positive view of the plan to relocate the capital of Indonesia.

In general, the results of this study show that both methods have the ability to conduct sentiment analysis on the relocation of the Indonesian capital. In choosing the right method, it is necessary to consider the purpose of the study, the type of data, and the level of complexity of the data used. In addition, the results of sentiment analysis can help the government in making the right decision on the move of the Indonesian capital city.

4. CONCLUSION

Based on the results of the research that has been done, it can be interpreted that sentiment analysis can be used to classify public sentiment towards the relocation of the Indonesian capital city. In this study, two classification methods were used, namely Naïve Bayes and Support Vector Machine. The test results show that SVM provides higher accuracy than Naïve Bayes, which is 0.897 and 0.802 respectively. In addition, it was also found that the majority of public sentiment towards the relocation of the national capital was positive, namely 56%. Neutral sentiment is 32%, and negative sentiment is 11%.

These results show that Indonesians tend to support moving the capital to Kalimantan. In this case, the government can use the results of sentiment analysis to improve policies and strategies in order to promote the plan to move the capital of Indonesia. In this study, the data used came from Twitter and YouTube comments between May 18 and July 6, 2022. Therefore, the results of this study only represent public sentiment at the time and media studied. The limitations of this research are the number of models and evaluations that are used only between Naive Bayes and support vector machines. However, it is hoped that the results of this study can provide a more comprehensive view and can be used as a reference in making decisions in the future. In conclusion, sentiment analysis using SVM can be a more effective alternative in classifying public sentiment towards moving the Indonesian capital city, and can provide a more accurate picture of people's views on the plan.

ACKNOWLEDGEMENTS

We would like to express our gratitude for the support and assistance provided by various parties in this research. First of all, we would like to thank the University Of Darussalam Gontor, which has provided the necessary facilities and resources during the research. We would also like to thank Aziz Mustafa and Taufiqurrahman, who provided invaluable direction and guidance in the process of this research. We acknowledge that this research would not have been successful without the support and contributions of all the parties mentioned above. We hope that the results of this research can provide benefits for the development of science and technology in the future. Once again, we thank you for the support and assistance that has been given during this research.

REFERENCES

Anggraini, N., Harahap, E.S.N. and Kurniawan, T.B. (2021) 'Text mining - Analisis teks terkait isu vaksinasi COVID-19', *Jurnal Ilmu Pengetahuan dan Teknologi Komunikasi*, 23(2), pp. 141–153. Available at: http://dx.doi.org/10.33169/iptekkom.23.2.2021.141-153.

Apriyani, H. (2020) 'Perbandingan Metode Naïve Bayes Dan Support Vector Machine Dalam Klasifikasi Penyakit Diabetes Melitus', 1(3), pp. 133–143.

Arsi, P., Kusuma, B.A. and Nurhakim, A. (2021) 'Analisis Sentimen Pindah Ibu Kota Berbasis Naive Bayes Classifier', *Jurnal Informatika Upgris*, 7(1), pp. 1–6. Available at: https://doi.org/10.26877/jiu.v7i1.7636.

Arsi, P. and Waluyo, R. (2021) 'Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM)', *Jurnal Teknologi Informasi dan Ilmu Komputer*, 8(1), p. 147. Available at: https://doi.org/10.25126/jtiik.0813944.

Atika, D., Styawati and Ari Aldino, A. (2022) 'Term Frequency-Inverse Document Frequency Support Vector Machine Untuk Analisis Sentimen Opini Masyarakat Terhadap Tekanan Mental Pada Media Sosial Twitter', *Jurnal Teknologi dan Sistem Informasi (JTSI)*, 3(4), p. page-page. Available at: http://jim.teknokrat.ac.id/index.php/JTSI.

Fitriyah, N., Warsito, B. and Maruddani, D.A.I. (2020) 'Analisis Sentimen Gojek Pada Media Sosial Twitter Dengan Klasifikasi Support Vector Machine (Svm', *Jurnal Gaussian*, 9(3), pp. 376–390. Available at: https://doi.org/10.14710/j.gauss.v9i3.28932.

Hidayat, T.F.T., Garno, G. and Ridha, A.A. (2021) 'Analisis Sentimen Opini Pemindahan Ibu Kota Pada Twitter Dengan Metode Support Vector Machine', *Jurnal Ilmu Komputer*, 14(1), p. 49. Available at: https://doi.org/10.24843/jik.2021.v14.i01.p06.

Laurensz, B. and Eko Sediyono (2021) 'Analisis Sentimen Masyarakat terhadap Tindakan Vaksinasi dalam Upaya Mengatasi Pandemi Covid-19', *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, 10(2), pp. 118–123. Available at: https://doi.org/10.22146/jnteti.v10i2.1421.

Lorena., S. (2016) 'Teknik Data Mining Menggunakan Metode Bayes Classifier Untuk Optimalisasi Pencarian Aplikasi Perpustakaan', *Jurnal Teknik Komputer*, 4(2), pp. 17–20.

Mas'udah, E., Wahyuni, E.D. and Arifiyanti, A.A. (2020) 'Analisis Sentimen: Pemindahan Ibu Kota Indonesia Pada Twitter', *Jurnal Informatika dan Sistem Informasi (JIFoSI)*, 1(2), pp. 397–401.

Monika Parapat, I. and Tanzil Furqon, M. (2018) 'Penerapan Metode Support Vector Machine (SVM) Pada Klasifikasi Penyimpangan Tumbuh Kembang Anak', *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2(10), pp. 3163–3169. Available at: http://j-ptiik.ub.ac.id.

Nanda, R. *et al.* (2022) 'Klasifikasi Berita Menggunakan Metode Support Vector Machine', *Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)*, 5(2), pp. 269–278. Available at: https://doi.org/10.32672/jnkti.v5i2.4193.

Nasution, M.R.A. and Hayaty, M. (2019) 'Perbandingan Akurasi dan Waktu Proses Algoritma K-NN dan SVM dalam Analisis Sentimen Twitter', *Jurnal Informatika*, 6(2), pp. 226–235. Available at: https://doi.org/10.31311/ji.v6i2.5129.

Nur Jamal Shaid (2022) *6 Alasan Ibu Kota Negara Pindah dari Jakarta ke Kalimantan Timur*, *Kompas.com*. Available at: https://money.kompas.com/read/2022/02/11/052456426/6-alasan-ibu-kota-negara-pindah-dari-jakarta-ke-kalimantan-timur?page=all (Accessed: 11 June 2022).

Sholihin, A. *et al.* (2019) 'Sains, Aplikasi, Komputasi dan Teknologi Informasi Analisis Penyakit Difteri Berbasis Twitter Menggunakan Algoritma Naïve Bayes', *Sakti*, 1(1), p. 7.

Wardani, N.S., Prahutama, A. and Kartikasari, P. (2020) 'Analisis Sentimen Pemindahan Ibu Kota Negara Dengan Klasifikasi Naïve Bayes Untuk Model Bernoulli Dan Multinomial', *Jurnal Gaussian*, 9(3), pp. 237–246. Available at: https://doi.org/10.14710/j.gauss.v9i3.27963.

Wibowo, P. (2016) *Implementasi Maqashid Syariah dalam Kepemimpinan Publik*, *4 Februari 2016*. Available at: https://www.kompasiana.com/pandu_wibowo/56b2cf532223bd5808ef6e07/implementasi-maqashid-syariah-dalam-kepemimpinan-publik.

Yulita, W. (2021) 'Analisis Sentimen Terhadap Opini Masyarakat Tentang Vaksin Covid-19 Menggunakan Algoritma Naïve Bayes Classifier', *Jurnal Data Mining dan Sistem Informasi*, 2(2), p. 1. Available at: https://doi.org/10.33365/jdmsi.v2i2.1344.