Terbit *online* pada laman web jurnal : http://innovatics.unsil.ac.id

Innovation in Research of Informatics (INNOVATICS)

| ISSN (Online) 2656-8993 |

Naive Bayes and Wordcloud for Sentiment Analysis of Halal Tourism in Lombok Island Indonesia

Irvandi[1], Bambang Irawan[2], Odi Nurdiawan[3]

[1,2,3]Teknik Informatika, STMIK IKMI Cirebon, Jl. Perjuangan No.10B, Kota Cirebon 45131, Indonesia
[1]irvandi2907@gmail.com, [2]bambang_irawan_2000@yahoo.com, [3]odinurdiawan2020@gmail.com

INFORMASI ARTIKEL

Article History:
Received: 19 February 2023
Last Revision: 20 May 2023
Published *Online*: 21 May 2023

KATA KUNCI

Sentiment Analysis,
Web Scraping,
Naive Bayes,
Word Cloud,
Google Maps

KORESPONDENSI

Telepon: +6289662299994

E-mail: *irvandi2907@gmail.com*

ABSTRACT

Lombok is one of the halal tourist destinations in Indonesia and has been recognized by the world. To examine these assumptions based on sources from tourist opinion, it is necessary to carry out a sentiment analysis whether their presence is as expected. Google Maps is a platform that can show the location of the island of Lombok along with written reviews from tourists who have visited. The collection of review data is done through the Web Scraping technique on the Google Maps Review, then the data is processed using RapidMiner. The algorithm used is Naive Bayes, an algorithm that uses probability or the concept of opportunity in classification. A word cloud visualization is also displayed to bring up words that tourists often say. 1493 data were obtained after Web scraping and cleansing had been labeled with positive and negative sentiment categories. Preprocessing is carried out which includes tokenize, filter token by length, transform case, stopword, and stemming, then classification using the Naive Bayes algorithm. From the results of testing the Naive Bayes algorithm model, an accuracy rate of 74.75 %. Word Cloud visualization also found the top words included "indah", "wisata". “pantai”, “alam”, “gunung”, and “masjid”.

1. INTRODUCTION

Indonesia is a country with the largest Muslim population in the world. According to a report from The Royal Islamic Strategic Studies Center (RISSC) there are 237.56 million Indonesians who are Muslim[1]. This can increase the potential of the tourism industry related to the existence of Indonesian Muslims in the eyes of the world by offering tourism services according to Islamic teachings, which today are called halal tourism. Halal tourism is tourism which in practice is permissible according to Islamic teachings, both accommodation, attractions and tourist objects [2]. Lombok Island previously occupied the top position as the best halal tourist destination in the world according to the 2019 GMTI (Global Muslim Travel Index) version.

Sentiment analysis is a field of study that analyzes written opinions which are used as a reference in making decisions in various situations and has the aim of analyzing opinions, judgments, and emotions related to a topic, product, and service [3][4]. Lombok Island has visitor review data on Google Maps which can be processed into information that can be used to summarize the opinions and sentiments of the majority of tourists on the island. The review feature on Google Maps is one of the things from the big data era where at this time everyone can put traces after they visit a place [5]. Classification algorithms can play a role in conducting sentiment analysis, one of which is using Naive Bayes which is an algorithm with the highest probability concept and can predict events based on classification results well [6] [7].

Related research that conducts sentiment analysis of tourist objects has been studied with Bali tourism objects [8] get the results from the implementation of the Naive Bayes algorithm to this study with a fairly good level of accuracy. Of the 5 tourist objects studied, Nusa Penida was obtained as a recommended tourist attraction because it has

an accuracy of 94.64 %. Garuda Wisnu Kencana with an accuracy value of 82.86 %, the edge with an accuracy value of 80%, Pandawa with an accuracy value of 90.71%, Pura Luhur Uluwutu with an accuracy value of 85.54%.

Halal tourism on Lombok Island is not yet clarified for new tourists, will its existence be as expected. So this study aims to uncover facts with data sourced from reviews of tourists who have visited the island of Lombok.

2. RELATED RESEARCH

Previous research that has been done as in research [9] discusses the implementation of the Naive Bayes algorithm for tourist sentiment which aims to be used as material for evaluating services and developing tourist objects in Sukabumi Regency. The review data from this study came from TripAdvisor and Google Maps with a total of 3,194 reviews which were used as a dataset after the crawling and labeling processes were carried out. The application of the naive Bayes classification method produces an accuracy of 78 %.

Then in paper [6] comparing the performance of naive bayes classification with lexicon based in sentiment analysis. From the research results, the comparison between the results of the classification accuracy carried out between the lexicon based method and the Naive Bayes classifier is 67 % and 78 %. With different accuracies, it shows that sentiment analysis using the naive Bayes classifier method has higher accuracy than the lexicon based method.

Based on the related research that has been described, it can be an illustration of researchers in conducting sentiment analysis research using the Naive Bayes algorithm implemented in this study.

3. METHODOLOGY

The steps of the method used in this study from start to finish are based on the following figure 1:

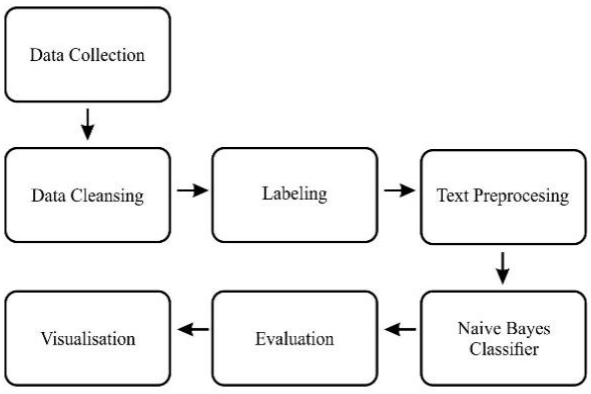

FIGURE 1. RESEARCH FLOW

3.1 Dataset

The dataset in the form of reviews from tourists was obtained from reviews on the Google Maps site. In preparing the dataset in this study, there were 3 stages that were passed including:

a) Data Collection

Data collection uses web scraping techniques on the Apify tool with only data taken in the form of review text which is then stored in an Excel file. Web scraping is a technique for getting information from websites automatically without having to copy it manually, focusing on getting data by extraction and extraction [10].

b) Data Cleansing

Data cleansing is the process of cleaning data from noise such as punctuation and other characters that are not important [11].

c) Labeling

The collected dataset was then manually labeled for each review with 2 categories, namely positive and negative.

3.2 Text Preprocesing

Unstructured data information will be analyzed in the process of searching for a certain pattern. In text preprocessing will go through several stages [8] between:

a) Transform Cases: the stage of equating all the letters in the data, in this process the letters will be converted to lowercase.
b) Tokenizing: the stage of separating the text based on the length of the text.
c) Filtering: the stage takes important words from the token results. can use the stoplist algorithm (discarding less important words) or wordlist (saving important words).
d) Stemming: the stage of removing affixes so that they become basic words.

3.3 Naive Bayes Clasifier

Naive Bayes is an algorithm that uses probability or the concept of opportunity in classification for sentiment analysis. Naive Bayes classification is also calculated in an easy-to-use and simple algorithm and can estimate an event based on the results of a good classification [6]. The following is the probability equation of the Naive Bayes method.

$$P(X|Y) = \frac{P(X|Y) \cdot P(X)}{P(Y)} \quad (1)$$

Which is:
X = Temporary estimates of data from a specific class
Y = Data with an unknown class
P(X|Y) = Estimated probability of X on condition Y (Posterior probability)
P(X) = Estimated probability of X (prior probability)
P(Y|X) = Estimated probability of Y with X
P(Y) = Opportunity Y

Posterior probability: there is a possibility of class X
Prior probability: the probability of the initial sample of class Y

4. EXPERIMENT AND RESULTS ANALYSIS

4.1 Data Collection

The tourist review dataset is shown from the coordinates of the island of Lombok on Google Maps reviews with a total number of 2700 reviews before being processed. Obtained by web scraping technique and stored in a Microsoft Excel file.

TABLE 1. WEB SCRAPING RESULT DATA

No.	Text
1	Saya ulas pengalaman saya selama beberapa kali menjelajahi berbagai spot indah di pulau ini. Snorkeling ...
2	Tempatx sejuk... Cocok untuk menenangkan diri
3	Perjalanan yg sangat indah trip Jawa Bali Lombok
4	Pulau Lombok memiliki panorama alam yang sangat cantik, ada berbagai destinasi wisata yang bisa anda kunjungi di pulau Lombok...
5	tempat wisata sekaligus ajang olahraga otomotif nasional dan internasional ..
..	..
2700	Pulau yang menyimpan banyak potensi wisata yang tak kalah jauh dari pulau bali...

4.2 Data Cleansing

The cleansing process is carried out on RapidMiner. Cleansing carried out in this process will remove unnecessary emoticons or characters, for example deleting hashtags, deleting punctuation characters, deleting non-ASCII characters, deleting foreign languages can also be done in this process because the analysis to be carried out focuses on Indonesian. just. Here are the steps:

1) Calling data stored in an excel file, then going through the model process as shown in figure 2.

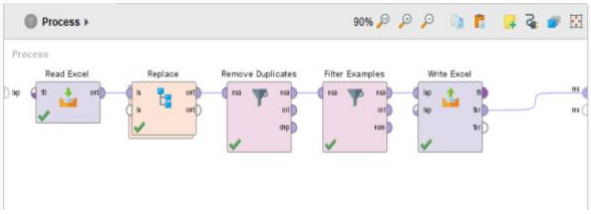

FIGURE 2. DATA CLEANSING MODEL PROCESS

2) In the replace subprocess section, there are several operators for deleting data that is not needed. Can be

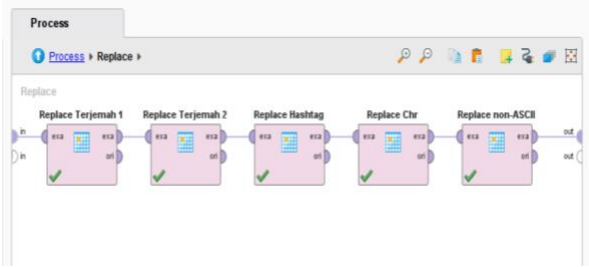

FIGURE 3. SUBPROCESS REPLACE OPERATOR

seen in figure 3.

With the following:

a) Replace Terjemah: to remove foreign languages before and after automatic translation when obtained from scraping results.

b) Replace Hashtag: to remove the hashtag followed by the following word with no space after the hashtag.

TABLE 2. REPLACE HASHTAG

Input	Output
#LetsGuide Aku bukan tipe yang suka berjemur dipantai. Pantai favorit juga hanya disekitaran Jawa Timur karena letak pantai dan gunung yang berdekatan, membuat pantai Jawa Timur lebih sejuk. Berbeda dengan Lombok, dia punya pesonanya sendiri. Lombok memang benar-benar tempat eksotis.	Aku bukan tipe yang suka berjemur dipantai. Pantai favorit juga hanya disekitaran Jawa Timur karena letak pantai dan gunung yang berdekatan, membuat pantai Jawa Timur lebih sejuk. Berbeda dengan Lombok, dia punya pesonanya sendiri. Lombok memang benar-benar tempat eksotis.

c) Replace Chr: to remove unnecessary punctuation characters from the dataset.

TABLE 3. REPLACE CHR

Input	Output
bagus dan bersih. nyaman dan tepat untuk di kunjungi?	bagus dan bersih nyaman dan tepat untuk di kunjungi
Telah terjadi tsunami di akibat kan gempa bumi.....!!!!!!	Pulau yang indah dan banyak tempat wisata
Lombok tanah kelahiran yg subur & indah	Lombok tanah kelahiran yg subur indah

d) Replace non-ASCII: to remove non-ASCII-based characters or symbols

TABLE 4. REPLACE NON-ASCII

Input	Output
Lombok bikin gak mau pulang ❤️❤️	Lombok bikin gak mau pulang

3) The final stage of the data cleansing process produces an excel file with a total of 1493 review data from the previous 2700 data.

4.3 Labeling

The process of labeling the data totaling 1493 was done manually with positive and negative sentiment categories, can be seen in table 5.

TABLE 5. LOMBOK ISLAND SENTIMENT DATASET

No.	Text	Sentiment
1	Perjalanan yg sangat indah trip Jawa Bali Lombok	positif
2	Pulau Lombok memiliki panorama alam yang sangat cantik ada berbagai destinasi wisata yang bisa anda kunjungi di pulau Lombok	positif
3	tempat wisata sekaligus ajang olahraga otomotif nasional dan internasional sebagai pulau masa depan utk kunjungan wisata lokal maupun internasional	positif
4	Lombok meneyediakan banyak tempat untuk destinasi liburan	positif
5	tempat wisata sekaligus ajang olahraga otomotif nasional dan internasional ..	
..	..	
1493	Pemandangan oke SDM kurang	negatif

4.4 Text Preprocesing

The stages of text preprocessing using RapidMiner can be seen in Figure 4. Which is a subprocess of the document from the data operator.

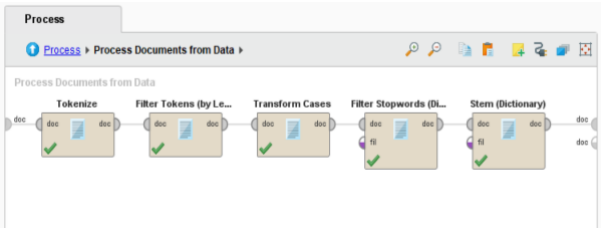

FIGURE 4. TEXT PREPROCESING MODEL

Which the following:

1.) Tokenize : separate words in a sentence

TABLE 6. TOKENIZE PROCESS

No.	Input	Output
1	Perjalanan yg sangat indah trip Jawa Bali Lombok	'Perjalanan' 'yg' 'sangat' 'indah' 'trip' 'Jawa' 'Bali' 'Lombok'
2	Pulau Lombok memiliki panorama alam yang sangat cantik ada berbagai destinasi wisata yang bisa anda kunjungi di pulau Lombok	'Pulau' 'Lombok' 'memiliki' 'panorama' 'alam' 'yang' 'sangat' 'cantik' 'ada' 'berbagai' 'destinasi' 'wisata' 'yang' 'bisa' 'anda' 'kunjungi' 'di' 'pulau' 'Lombok'

2.) Filter tokens (By length) : conditions the word to be filtered to only have a certain character length with a minimum of 4 characters and a maximum of 22 characters.

TABLE 7. FILTER TOKEN BY LENGTH PROCESS

No.	Input	Output
1	'Perjalanan' 'yg' 'sangat' 'indah' 'trip' 'Jawa' 'Bali' 'Lombok'	'Perjalanan' 'sangat' 'indah' 'trip' 'Jawa' 'Bali' 'Lombok'
2	'Pulau' 'Lombok' 'memiliki' 'panorama' 'alam' 'yang' 'sangat' 'cantik' 'ada' 'berbagai' 'destinasi' 'wisata' 'yang' 'bisa' 'anda' 'kunjungi' 'di' 'pulau' 'Lombok'	'Pulau' 'Lombok' 'memiliki' 'panorama' 'alam' 'yang' 'sangat' 'cantik' 'berbagai' 'destinasi' 'wisata' 'yang' 'bisa' 'anda' 'kunjungi' 'pulau' 'Lombok'

3.) Transform Cases : converts all types of letters to lowercase.

TABLE 8. TRANFORM CASES PROCESS

No.	Input	Output
1	'Perjalanan' 'sangat' 'indah' 'trip' 'Jawa' 'Bali' 'Lombok'	'perjalanan' 'sangat' 'indah' 'trip' 'jawa' 'bali' 'lombok'
2	'Pulau' 'Lombok' 'memiliki' 'panorama' 'alam' 'yang' 'sangat' 'cantik' 'berbagai' 'destinasi' 'wisata' 'yang' 'bisa' 'anda' 'kunjungi' 'pulau' 'Lombok'	'pulau' 'lombok' 'memiliki' 'panorama' 'alam' 'yang' 'sangat' 'cantik' 'berbagai' 'destinasi' 'wisata' 'yang' 'bisa' 'anda' 'kunjungi' 'pulau' 'lombok'

4.) Stopword Removal (filtering stopword): delete certain words that are not needed. This stopword removal uses an Indonesian dictionary sourced from the Kaggle site which adds relevant stopword words based on the researched data.

TABLE 9. STOPWORD REMOVAL PROCESS

No.	Input	Output
1	'perjalanan' 'sangat' 'indah' 'trip' 'jawa' 'bali' 'lombok'	'perjalanan' 'indah' 'trip' 'jawa' 'bali' 'lombok'
2	'pulau' 'lombok' 'memiliki' 'panorama' 'alam' 'yang' 'sangat' 'cantik' 'berbagai' 'destinasi' 'wisata' 'yang' 'bisa' 'anda' 'kunjungi' 'pulau' 'lombok'	'pulau' 'lombok' 'panorama' 'alam' 'cantik' 'destinasi' 'wisata' 'pulau' 'lombok'

5.) Stemming: is the process of changing affixes into root words. Stemming uses an Indonesian dictionary created by researchers based on a dataset with the highest frequency of affixed words appearing.

TABLE 10. STEMMING PROCESS

No.	Input	Output
1	'perjalanan' 'indah' 'trip' 'jawa' 'bali' 'lombok'	'jalan' 'indah' 'trip' 'jawa' 'bali' 'lombok'

4.5 Building Naive Bayes Clasifier

Classification analysis using the Naive Bayes algorithm using RapidMiner, the operator of which has been adjusted can be seen in Figure 5.

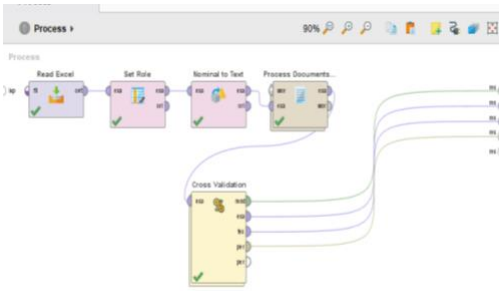

FIGURE 5. NAIVE BAYES MODEL DESIGN

Which the following:

1.) Read Excel: reading the data with the xlsx file extension which is a dataset resulting from data cleansing processing that has been given positive and negative sentiment.
2.) Set Role: make the sentiment column to be used as a label in the processing of the Naive Bayes algorithm.
3.) Nominal to Text: change the nominal data type to text which will be forwarded to the next operator.
4.) Process Documents from Data: this operator converts the text data used in the TF-IDF weighted, has a subprocess that is filled as the text preprocessing.
5.) Cross Validation: is a validation technique in which the process will be divided into training data and testing data randomly in K parts, using an iteration called k-fold validation, with a value of K = 10. In the cross

validation subprocess, the contents of the Naive Bayes algorithm are in the training column, then in the the testing column contains the apply model and performance which can be seen in Figure 6.

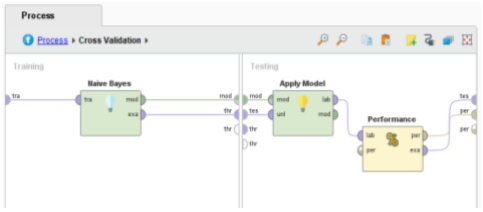

FIGURE 6. CROSS VALIDATION

4.6 Evaluation

Validation using the value of K = 10, produces a confusion matrix to get an evaluation. The results of the accuracy obtained were 74.75% of the 1493 review data that were used as datasets. This means that the system can classify Lombok Island sentiment into positive and

accuracy: 74.75% +/- 5.21% (micro average: 74.75%)

	true positif	true negatif	class precision
pred. positif	1085	24	97.84%
pred. negatif	353	31	8.07%
class recall	75.45%	56.36%	

FIGURE 7. CONFUSSION MATRIX

From Figure 7, the quality of the categories obtained can be calculated using the formula below:
TP (True Positif) = 1085
TN (True Negatif) = 31
FP (False Positif) = 24
FN (False Negatif) = 353

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{1085+31}{1085+31+24+353} = \frac{1116}{1493} = 0{,}7475 = 74{,}75\%$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{1085}{1085+353} = \frac{1085}{1438} = 0{,}7545 = 75{,}45\%$$

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{1085}{1085+24} = \frac{1085}{1109} = 0{,}9784 = 97{,}84\%$$

4.7 Visualization

Wordcloud visualization is raised to get any opinions that are often expressed by tourists visiting the island of Lombok. With regard to halal tourism, it is certain that the island of Lombok has several qualified facilities to provide services for tourists, especially those who are Muslim, one of which is the existence of mosque facilities that are easy to find. Proven by word "masjid" in the Wordcloud visualization on the 25 highest word frequencies overall. Then the situation on the island of Lombok seems comfortable, as evidenced by "nyaman" and the residents are polite called "ramah" to tourists, and has beautiful natural scenery of the beach and mountains.

FIGURE 8. WORDCLOUD

5. CONCLUCION

Based on the results of the classification by the Naive Bayes algorithm to carry out a sentiment analysis for halal tourism on Lombok island which is divided into 2 categories of positive and negative sentiments with data obtained from Lombok island review data on Google Maps showing an accuracy of 74.75%, recall 75.45%, and precision 97.84%. The majority of tourist opinions are proven in the Wordcloud visualization which has the highest frequency of words especially "indah", "wisata". "pantai", "alam", "gunung", and "masjid" which can be a reference as a description of the island of Lombok. The data taken is general, so that it becomes a correlation with halal tourism only mosque called "masjid".

REFERENCES

[1] M. Ayu Rizaty, "Jumlah Penduduk Muslim Indonesia Terbesar di Dunia pada 2022," *DataIndonesia.id*, Nov. 03, 2022. https://dataindonesia.id/ragam/detail/populasi-muslim-indonesia-terbesar-di-dunia-pada-2022 (accessed Jan. 09, 2023).

[2] S. R. Pratiwi, S. Dida, and N. A. Sjafirah, "Strategi Komunikasi dalam Membangun Awareness Wisata Halal di Kota Bandung," *Jurnal Kajian Komunikasi*, vol. 6, no. 1, 2018, doi: 10.24198/jkk.v6i1.12985.

[3] V. W. D. Thomas and F. Rumaisa, "Analisis Sentimen Ulasan Hotel Bahasa Indonesia Menggunakan Support Vector Machine dan TF-IDF," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 6, no. 3, p. 1767, Jul. 2022, doi: 10.30865/mib.v6i3.4218.

[4] A. Rifa'i, H. Sujaini, and D. Prawira, "Sentiment Analysis Objek Wisata Kalimantan Barat Pada Google Maps Menggunakan Metode Naive Bayes," *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, vol. 7, no. 3, p. 400, Dec. 2021, doi: 10.26418/jp.v7i3.48132.

[5] F. U. Haq, "PENGGUNAAN GOOGLE REVIEW SEBAGAI PENILAIAN KEPUASAN PENGUNJUNG DALAM PARIWISATA," *Tornare*, vol. 2, no. 1, p. 10, Jan. 2020, doi: 10.24198/tornare.v2i1.25826.

[6] F. Amaliah and D. I. K. Nuryana, "Perbandingan Akurasi Metode Lexicon Based Dan Naive Bayes

Classifier Pada Analisis Sentimen Pendapat Masyarakat Terhadap Aplikasi Investasi Pada Media Twitter," *Journal of Informatics and Computer Science*, vol. 03, 2022.

[7] B. S. Gandhi, D. A. Megawaty, and D. Alita, "Aplikasi Monitoring dan Penentuan Peringkat Kelas Menggunakan Naive Bayes Classifier," *Jurnal Informatika dan Rekayasa Perangkat Lunak*, vol. 2, no. 1, pp. 54–63, Apr. 2021, doi: 10.33365/jatika.v2i1.722.

[8] D. Siti Utami and A. Erfina, "Analisis Sentimen Objek Wisata Bali Di Google Maps Menggunakan Algoritma Naive Bayes," 2022.

[9] B. R. Atmadja, "Analisis Sentimen Bahasa Indonesia Pada Tempat Wisata di Kabupaten Sukabumi Dengan Naïve Bayes," vol. 15, no. 2, pp. 371–382, 2022, [Online]. Available: http://journal.stekom.ac.id/index.php/elkom□page371

[10] D. A. Deviacita, H. P. Sasty, and H. Muhardi, "Implementasi Web Scraping untuk Pengambilan Data pada Situs Marketplace," vol. 7, no. 4, 2019.

[11] Sudianto, P. Wahyuningtias, H. Warih Utami, U. Ahda Raihan, H. Nur Hanifah, and Y. Nicholas Adanson, "COMPARISON OF RANDOM FOREST AND SUPPORT VECTOR MACHINE METHODS ON TWITTER SENTIMENT ANALYSIS (CASE STUDY: INTERNET SELEBGRAM RACHEL VENNYA ESCAPE FROM QUARANTINE)," *Jurnal Teknik Informatika (JUTIF)*, vol. 3, no. 1, pp. 141–145, 2022, doi: 10.20884/1.jutif.2022.3.1.168.

BIODATA OF AUTHORS

Irvandi
Born in Cirebon, July 29, 1999. STMIK IKMI Cirebon Students at Informatics Engineering Study Program.

Bambang Irawan, M.T
Lecturer of entrepreneurship courses at STMIK IKMI Cirebon in the Informatics Engineering Study Program.

Odi Nurdiawan, M.Kom
Born in Indramayu, April 12, 1991. Currently a Lecturer at STMIK IKMI Cirebon. Bachelor of Informatics Engineering Study at the Islamic University of Indonesia, Yogyakarta, graduated in 2014; S2 Information Systems STMIK LIKMI, Bandung, graduated in 2017.