ISSN 2085-4552

Sentiment Analysis of An Internet Provider Company Based on Twitter Using Support Vector Machine and Naïve Bayes Method

Farhan Hashfi[1], Dedy Sugiarto[2], Is Mardianto[3]
[1,3] Program Studi Teknik Informatika, Fakultas Teknologi Industri, Universitas Trisakti
[2] Program Studi Sistem Informasi, Fakultas Teknologi Industri, Universitas Trisakti
[1]farhan0640017000008@trisakti.ac.id, [2]dedy@trisakti.ac.id, [3]mardianto@trisakti.ac.id

Accepted 12 December 2021
Approved 20 February 2022

***Abstract*— Tweets from users in the form of opinions about a product can be used as a company evaluation of the product. To obtain this evaluation, the method that can be used is sentiment analysis to divide opinions into positive and negative opinions. This study uses 1000 data from Twitter related to an internet service provider company where the data is divided into two classes, namely 692 positive classes and 308 negative classes. In the Tweet there are still many words that are not standard. Therefore, previously carried out the initial process or preprocessing to filter out non-standard words. Before doing the classification, the data needs to be divided into training data and test data with a ratio of 90:10, then processed using the Support Vector Machine and Naïve Bayes techniques to get the results of the classification of positive opinions and negative opinions. The level of accuracy in the classification using the Support Vector Machine is 84% and using Naïve Bayes is 82%.**

***Index Terms*— Internet Provider; Naïve Bayes; Sentiment Analysis; Support Vector Machine; Twitter**.

I. INTRODUCTION

The internet has developed very rapidly to date in influencing media and communication. One of the factors supporting the success of the Internet in Indonesia is that infrastructure development has reached remote areas in Indonesia [1]. This can be proven by the increasing use of social media. Social Media is an Internet service most commonly used by Indonesian citizens. One of them is Twitter.

Twitter is used for various things such as sharing personal things, using it to sell, to reporting an opinion to a brand or company. Information shared on Twitter is typically 140 characters long [2]. In general, a company uses social media to gather information about the goods or services they offer. The most common use of social media by companies is to use social media for marketing activities and social media for customer service [3]. Therefore, the opinion group of Twitter users will be influenced by the emotions (emotions) that are classified in order to determine their polarization, namely positive opinions or negative opinions.

Sentiment analysis is the process of using text analysis to derive various data sources from the Internet and various other social media platforms. One of the purposes of sentiment analysis is to get someone's opinion on a company service and then classify that opinion into positive opinions and negative sentiments [4]. In conducting sentiment analysis, the technique used to retrieve data from Twitter uses the Crawling technique which requires API permission from the platform itself. Furthermore, the techniques for classifying the tweet data are Support Vector Machine, KNearest Neighbor, and Naïve Bayes [5]. By using this method, it produces a classification between two categories, namely positive opinions and negative opinions, where the results can be useful for observers of the company's data to determine the next marketing strategy.

II. THEORY

A. Sentiment Analysis

Sentiment analysis is one of the methods used to identify an opinion or sentiment expressed using a text or document and how that opinion is categorized as positive opinion and negative opinion. Basically, sentiment analysis tries to assess a different aspect of the standard language in order to help an agency or company to get positive opinions as well as negative opinions about the products they offer [6]. Sentiment itself can be interpreted as an emerging concept in which everyone's different emotions are determined from the content of the text, so that it can be processed to extract the opinions and sentiments of many people.

In sentiment analysis, there are 3 opinions that can be a reference for agencies or companies to obtain information on the quality of the products offered, namely positive opinions, negative opinions, and neutral opinions [7]. Sentiment refers to several topics, opinions on certain topics have different meanings from other opinions that are the same on other subjects. Analyst sentiment is usually used to determine the quality of services or the quality of a product from an

agency or company that wants to develop the services or products offered [8].

Sentiment analysis is a new part of research in Natural Language Processing (NLP) which aims to find subjectivity in texts or documents to classify opinions or sentiments. In the sentiment classification procedure, there are three techniques that can be used, namely Machine Learning, Lexicon Based, and Hybrid Approach. At this time, sentiment analysis is widely used using the Machine Learning method because this method is closer to the prediction of sentiment polarity based on the data that has been prepared [6].

B. *Data Crawling*

Crawling is a method of collecting data from a source to be analyzed or processed. Crawling is the first stage that is usually used to analyze the sentiments of social media users towards a product or service. Crawling can also be interpreted as a method of quickly gathering large numbers of web pages into a local storage and indexing them based on specified keywords [8]. Sentiment analysis usually uses the Crawling method on Twitter social media by utilizing the features provided by Twitter, namely Application Programming Interface Systems (APIs) [9].

Twitter provides features APIs that can be utilized using the crawling method to generate a collection of text data based on the desired keywords. Crawling tweet data on twitter is a method to get tweet data from twitter that gets data from the twitter server by utilizing these APIs features in the form of username data and tweet data as needed [9].

The data crawling method on Twitter is mostly done using the python programming language by utilizing the tweepy library provided by the python. That way, twitter data collection will be easier to obtain using the help of the python library and the APIs feature of twitter [10].

C. *Preprocessing*

Preprocessing is the stage of sentiment analysis to make documents more structured so that they are ready for analysis [7]. The steps taken for the preprocessing of this research are Case Folding, Cleansing, Tokenizing, Normalization, and Stopword Removal. In processing text mining data, it is necessary to carry out this stage [7].

1. *Case Folding*

   One of the steps of the text pre-processing stage which serves to convert all letters in the document to lowercase. This step is done to make the search easier.

2. *Cleansing*

   Cleansing is one of the steps used to remove links, mentions, hashtags, URLs, punctuation marks, numbers, and one letter. The purpose of this stage is to make the data used tidier for the next stage of text preprocessing.

3. *Tokenizing*

   This stage is used to make sentences or documents into words or tokens. This stage needs to be carried out for the next stage, namely changing all words that have wrong writing and are abbreviated. This stage is also carried out to carry out the word weighting stage.

4. *Normalization*

   After carrying out the Tokenizing stage, the next stage to do is the normalization stage, where this stage is used to change all words that have incorrect writing and are shortened to the words they should be. This stage is used so that there is no misunderstanding of the meaning of each word.

5. *Stopword Removal*

   The next step used for the text preprocessing process in this research is stopword removal. This stage is used to delete words that do not have important meaning.

D. *Term Frequency – Inverse Document Frequency (TF-IDF)*

Term Frequency-Inverse Document Frequency (TF-IDF) is a process for analyzing the interaction between phrases or sentences with multiple documents. The term frequency-reverse document frequency (TF-IDF) is the number of words available in a document. At the same time, the TF-IDF metric is statistical data that shows the importance of words in a data set or document [8]. Term frequency (TF) is an aspect that decides the weight of a word in a document based on the number of its presence in the document. When assigning weights to a word, the value of the number of occurrences of the word (term frequency) must be considered. The greater the weight given to the word, the greater the weight on a document, or provides a greater application value.

Inverse document frequency (IDF) is one aspect that reduces the excess of terms that often appear in documents. This is needed because terms that often appear in documents can be considered as general terms, so these terms are considered unimportant [5]. On the other hand, the term "scarcity factor in document collection" should be considered when determining the weights. TF-IDF can be disclosed namely:

$$TF - IDFt, d = TFt, d \times IDFt \quad (1)$$

Where IDFt = log N/DFt

t = counted words,
d = sentence weight (d),
TF-IDFtd = sentence weight (d) against word (t),

Tftd = Term Frequency,
IDFt = Inverse Document Frequency,
N = number of sentences,
DFt = number of words repeated

E. Support Vector Machine

A method used for linear classification is the Support Vector Machine (SVM), which can approximate the best ignition path, such as the separator between two categories. The supporting vector machine uses a linear function hypothesis space in the operation of high-dimensional features. The basic principle is linear classification, then it was developed to solve nonlinear problems by adding the concept of kernel techniques to high-dimensional workspaces [11].

$$a(B) = c\mathcal{x}\emptyset(B)+d \quad (2)$$

That is, B is a feature vector, c is a vector of different weights, is a function of non-linear mapping, and attribute d is a vector.

F. Naïve Bayes

Naïve Bayes is a classification method which holds that each term occurs as independent. In general, Naïve Bayes approach to probability or probability. The Naïve Bayes algorithm basically predicts future probabilities from past experiences [12].

$$p(C_k \mid x) = \frac{p(C_k)p(x|C_k)}{p(x)} \quad (3)$$

The above formula explains where P(x|C_k) is a conditional probability of the word x occurring in documents of class (C_k), P(C_k) is the previous probability of the dataset that has occurred in class C_k. P(x|ck) and P(ck) are estimated from the available data.

G. Evaluation

Evaluation was conducted to determine the accuracy of the modeling that has been applied to both methods. Then compare the results of two different data sets by applying the confusion matrix method. Confusion Matrix is a process that is generally used in data mining to calculate the level of accuracy. A classification system produces classification results and information will be loaded about the classification that has been correctly predicted by the Confusion Matrix [12]. Accuracy, Precision, and Recall are parameters to test the performance of calculations that have been done. Precision (P) is a parameter to find out how many results of processing that are relevant to the information you want to search for or with other bonds, namely the positive classification that is true (true positive) and the overall data which is predicted to be i-positive class. Precision can be obtained using equation [5].

$$P = \frac{tp}{tp+fp} \quad (4)$$

Recall (R) is how many irrelevant documents in the collection generated by the system or with other bindings are the number of documents which have a true classification positive (true positive) and all documents including true negative (true positive ones). Recall can be obtained by using equation [5].

$$R = \frac{tp}{tp+fn} \quad (5)$$

Furthermore, Accuracy (A) is the number of documents that are classified correctly, either true positive or true negative. Accuracy can be obtained by using equation [12].

$$A = \frac{(tp+fn)}{(tp+fp+tn+fn)} \quad (6)$$

Variables such as TN, TP, FN, and FP are derived from the confusion matrix [5].

TN = True Negative
TP = True Positive
FN = False Negative
FP = False Positive

III. METHOD

In this analysis, there are steps that are followed, namely generating tweet data from twitter using the twitter data crawling method, then the data is processed into more structured data using the preprocessing method, then the data classification uses the Naïve Bayes classification algorithm and the Support Vector Machine.

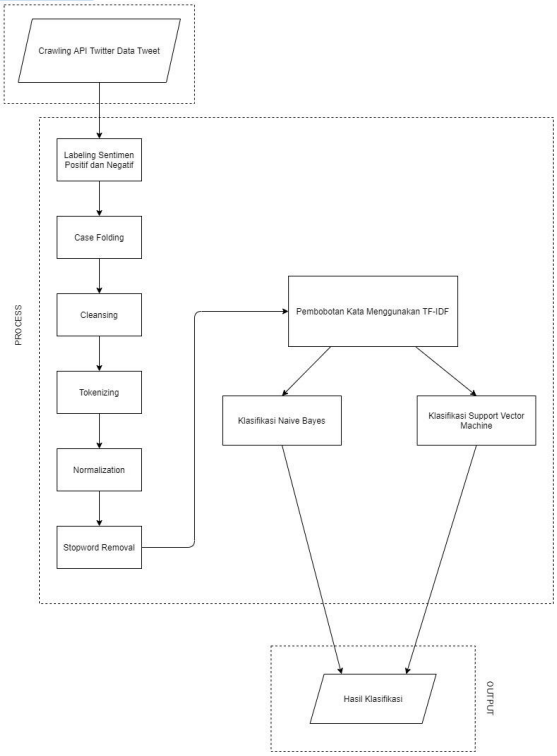

Fig 1. Flow of the Great Research Process

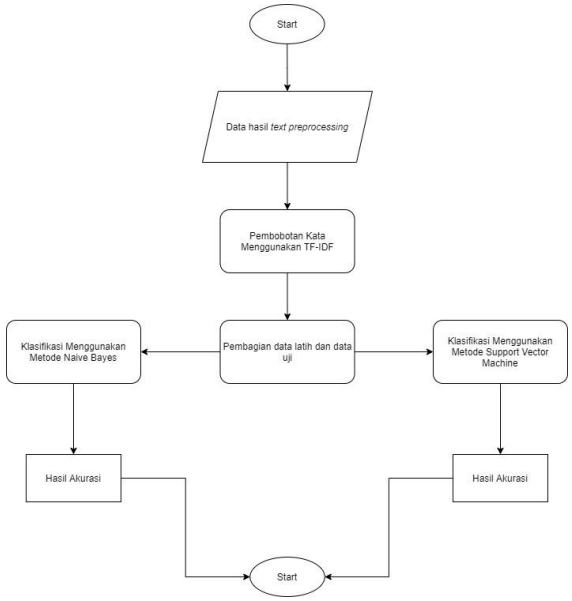

Fig 2. Classification Process Flow

Figure 1 shows the large process flow that was carried out to complete this research. Figure 2 shows the classification flow after the data has been structured through the text preprocessing process, involving two classification algorithms Naïve Bayes and Support Vector Machine [13].

IV. RESULT

The analysis process is carried out by testing and analyzing based on the results of sentiment analysis of Indonesian-speaking Indihome users on the Twitter platform using the Naive Bayes classification algorithm and Support Vector Machine [14].

A. Data

The tests carried out in this analysis used 1000 tweet data from Twitter with the keyword 'indihome' in Indonesian. This test uses 2 classes, namely positive and negative classes. The data collection is carried out using the Crawling technique that utilizes the tweepy library from the Python programming language. After that, labeling positive and negative sentiments on the dataset is done manually. For manual labeling of positive and negative sentiments, there should be the help of linguists in determining the positive or negative sentiments of an opinion. The author only uses two classes of sentiment, namely positive and negative because the author wants a more conical conclusion between the two classes.

TABLE I. TOTAL DATASET LABELS

Total Dataset Labels	
Positive	692
Negative	308
Total	1000

B. Text Preprocessing

After getting the dataset from the Crawling results and labeling positive and negative sentiments, the dataset needs to be changed into more neat and structured data to be processed using the Naïve Bayes algorithm and Support Vector Machine. Example of Text Preprocessing output based on the example of Table 2.

TABLE II. EXAMPLE OF TEXT PREPROCESSING

Before	After
indihome knp LOS dari tadi pagi?? ku mau drakoran help @IndiHomeCare	indihome kenapa los dari tadi pagi aku mau drakoran help

The Text Preprocessing step is needed to prepare the data so that the dataset becomes neater and more structured. Text Preprocessing has several stages in it, namely Case Folding, Cleansing, Tokenizing, Normalization, and Stopword Removal.

C. TF-IDF

After the dataset goes through the text preprocessing stage, then the dataset through the TF-IDF method functions for word weighting. Word weighting is needed to process datasets using the Naïve Bayes algorithm and Support Vector Machine. Example of term weighting based on Table 3.

TABLE III. TF-IDF WORD WEIGHTING EXAMPLE

Term	Weight
indihome	0.087980
banget	0.028923
rumah	0.026522
lambat	0.026136
kasih	0.022233

Word weighting utilizes the Scikit-learn library found in the Python programming language. The example in Table 3 shows 5 words that often appear in the dataset [15].

D. Classification Method

In this analysis, the classification process is carried out using two Naïve Bayes classification algorithms and the Support Vector Machine. From the results of the two classification methods have different accuracy values. The classification process is carried out using the Scikit-learn library contained in the Python programming language using the Naïve Bayes algorithm and the Support Vector Machine.

Before carrying out the classification method, the dataset needs to be divided into two parts including training data and test data. The ratio used in this study

uses 90:10 because according to research that has been done previously, it can be concluded that the higher the value of the training data, the better the level of accuracy obtained in the method used. The training data serves to train the model to recognize the existing patterns in the dataset, while the test data is used as test data for the classification method. In this study, the training data contained 90% of the dataset and the test data contained as much as 10% of the dataset. The results of testing the Naïve Bayes classification algorithm and Support Vector Machine are shown in Figure 3 and Figure 4

```
Multinomial Naive Bayes Acuracy: 82.0
Multinomial Naive Bayes Precision: 80.51948051948052
Multinomial Naive Bayes Recall: 95.38461538461539
Multinomial Naive Bayes f1_score: 87.32394366197184
==================================================
```

	precision	recall	f1-score	support
Negatif	0.81	0.95	0.87	65
Positif	0.87	0.57	0.69	35
accuracy			0.82	100
macro avg	0.84	0.76	0.78	100
weighted avg	0.83	0.82	0.81	100

Figure 3. Results of the Naïve Bayes method

```
Support Vector Machine Acuracy: 84.0
Support Vector Machine Precision: 82.66666666666667
Support Vector Machine Recall: 95.38461538461539
Support Vector Machine f1_score: 88.57142857142857
==================================================
```

	precision	recall	f1-score	support
Negatif	0.83	0.95	0.89	65
Positif	0.88	0.63	0.73	35
accuracy			0.84	100
macro avg	0.85	0.79	0.81	100
weighted avg	0.85	0.84	0.83	100

Figure 4. Results of the Support Vector Machine Method

The figure shows the results of Accuracy, Precision, Recall, and F1 Score from the two different classification methods, it can be seen that the accuracy generated in the Support Vector Machine calculation is higher than Naïve Bayes, namely the Accuracy value in the Support Vector Machine method is 84% while the accuracy value in the Support Vector Machine method is 84%. Naïve Bayes method is 82%.

E. Model Evaluation

The evaluation process uses a Confucian Matrix table to evaluate the results of the Naïve Bayes classification algorithm and Support Vector Machine. The results of the model evaluation can be seen based on Figure 5 and Figure 6

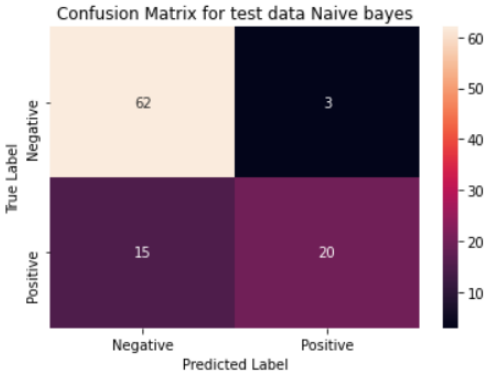

Fig 5. Evaluation of Naïve Bayes

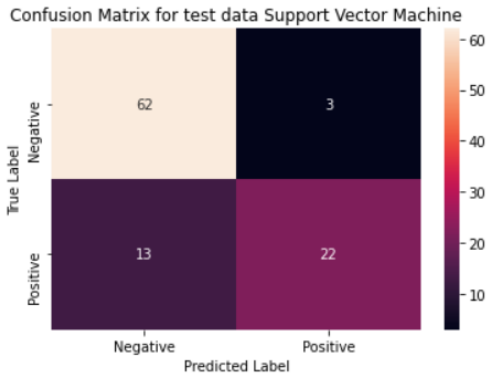

Fig 6. Evaluation of Support Vector Machine

From the evaluation value of the image above, it shows that the two classification methods carried out on this research dataset still have errors in classifying positive and negative sentiments, where both classification methods still place the positive sentiment class on negative sentiment, and vice versa.

V. CONCLUSIONS

In the results of research calculations that have been completed related to Indihome customer sentiment on Indihome services using the Naïve Bayes classification algorithm and Support Vector Machine to get the accuracy value, namely the accuracy of the Support Vector Machine algorithm is greater than the Naïve Bayes classification method. For this reason, in this study using 1000 Indihome customer datasets on the Twitter social media platform, the Support Vector Machine method is a better method than the Naïve Bayes method. Data is collected for 3 months starting from February 2021 to April 2021.

However, this research still has several shortcomings, namely the process of labeling positive and negative sentiments is done manually which produces more negative sentiments than positive sentiments. There are differences from the data labeling that is applied manually to test the model using the class prediction results from the model classification results. In addition, this study only uses 1000 datasets. The

accuracy of the Naive Bayes method is 82% while the Support Vector Machine is 84%.

REFERENCES

[1] F. Amalia, R. T. Sulistyo, and A. H. Brata, "Analisis Tingkat Penerimaan E-Learning Sebagai Alternatif Media Pembelajaran Pada Siswa SMK," *Smatika J.*, vol. 10, no. 02, pp. 41–47, 2020, doi: 10.32664/smatika.v10i02.450.

[2] A. W. Attabi, L. Muflikhah, and M. A. Fauzi, "Penerapan Analisis Sentimen untuk Menilai Suatu Produk pada Twitter Berbahasa Indonesia dengan Metode Naïve Bayes Classifier dan Information Gain," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 11, pp. 4548–4554, 2018.

[3] T. Yulianita, T. W. Utami, and M. Al Haris, "Analisis sentimen dalam penanganan covid-19 di indonesia menggunakan naive bayes classifier," *Semin. Nas. Variansi*, pp. 235–243, 2020.

[4] U. Rofiqoh, R. S. Perdana, and M. A. Fauzi, "Analisis Sentimen Tingkat Kepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonesia Pada Twitter Dengan Metode Support Vector Machine dan Lexion Based Feature," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 1, no. 12, pp. 1725–1732, 2017, [Online]. Available: http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/628.

[5] B. Gunawan, H. S. Pratiwi, and E. E. Pratama, "Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes," *J. Edukasi dan Penelit. Inform.*, vol. 4, no. 2, p. 113, 2018, doi: 10.26418/jp.v4i2.27526.

[6] H. Tuhuteru and A. Iriani, "Analisis Sentimen Perusahaan Listrik Negara Cabang Ambon Menggunakan Metode Support Vector Machine dan Naive Bayes Classifier," *J. Inform. J. Pengemb. IT*, vol. 3, no. 3, pp. 394–401, 2018, doi: 10.30591/jpit.v3i3.977.

[7] M. S. Hadna, P. I. Santosa, and W. W. Winarno, "Studi Literatur Tentang Perbandingan Metode Untuk Proses Analisis Sentimen Di Twitter," *Semin. Nas. Teknol. Inf. dan Komun.*, vol. 2016, no. Sentika, pp. 57–64, 2016, [Online]. Available: https://fti.uajy.ac.id/sentika/publikasi/makalah/2016/95.pdf.

[8] J. A. Septian, T. M. Fahrudin, and A. Nugroho, "Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF - IDF dan K - Nearest Neighbor," *J. Intell. Syst. Comput.*, no. September, pp. 43–49, 2019.

[9] E. S. Negara, R. Andryani, and P. H. Saksono, "Analisis Data Twitter: Ekstraksi dan Analisis Data G eospasial," *J. INKOM*, vol. 10, no. 1, p. 27, 2016, doi: 10.14203/j.inkom.433.

[10] F. A. Suharno and L. Listiyoko, "Aplikasi Berbasis Web dengan Metode Crawling sebagai Cara Pengumpulan Data untuk Mengambil Keputusan," *Semin. Nas. Rekayasa Teknol. Inf.*, no. November, pp. 105–109, 2018.

[11] A. A. Lutfi, A. E. Permanasari, and S. Fauziati, "Corrigendum: Sentiment Analysis in the Sales Review of Indonesian Marketplace by Utilizing Support Vector Machine," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 4, no. 2, p. 169, 2018, doi: 10.20473/jisebi.4.2.169.

[12] P. Antinasari, R. S. Perdana, and M. A. Fauzi, "Analisis Sentimen Tentang Opini Film Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naive Bayes Dengan Perbaikan Kata Tidak Baku," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 12, pp. 1733–1741, 2017, [Online]. Available: http://j-ptiik.ub.ac.id.

[13] P. Arsi and R. Waluyo, "Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM)," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 1, p. 147, 2021, doi: 10.25126/jtiik.0813944.

[14] E. Fitri, "Analisis Sentimen Terhadap Aplikasi Ruangguru Menggunakan Algoritma Naive Bayes, Random Forest Dan Support Vector Machine," *J. Transform.*, vol. 18, no. 1, p. 71, 2020, doi: 10.26623/transformatika.v18i1.2317.

[15] A. Sari, F. V., & Wibowo, "Analisis Sentimen Pelanggan Toko Online Jd. Id Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 2, no. 2, pp. 681–686, 2019.