




Information System Based on Sentiment Analysis: Study of the Application of Naive Bayes Classifier to Service Provider 3 on Twitter

Akhmad Chanafi^{1*}  Muhammad Arifin²  Arif Setiawan³ 

^{1,2,3} Information Systems Program, Faculty of Engineering, Universitas Muria Kudus, Kudus 59327, Indonesia Corresponding

Author Email: 201853045@std.umk.ac.id

Copyright: ©2025 The author(s). This article is published and is licensed under Information Systems Department Faculty of Engineering Universitas Muria Kudus (<https://jurnal.umk.ac.id/index.php/insytech>).

<https://doi.org/10.24176/insytech.v1i2.14606>

ABSTRACT

Received: January 14, 2025

Revised: January 19, 2025

Accepted: January 23, 2025

Available online: February 01, 2025

Keywords:

Sentiment analysis, Twitter, TriIndonesia, Naïve Bayes Classifier, SMOTE.

Twitter is used to exchange information, opinions on a topic circulating in the community. Of course, this can be used to find out what the public thinks about a product or hot news. With the Twitter social media, the information obtained is very diverse through tweets, the tweets themselves are written information in the form of raw data that can be processed into sentiment analysis. The data obtained and processed in this study is Twitter data with the keyword triIndonesia. The data will be divided into training data and test data and classified into 3 classes namely positive, neutral and negative using the Naïve Bayes classifier method. In the implementation of the test, the use of SMOTE is needed to overcome data imbalance, because the data obtained is not balanced. After going through 3 tests with different distribution of data sets, the highest accuracy value was obtained at 79% in the distribution of 90% training data and 10% testing data.

1. INTRODUCTION

According to data released in February 2023, the number of Twitter users in Indonesia reached around 24 million people[1]. Twitter used by many people to express their opinions and perceptions regarding a topic, including topics about the products or services they use. Provider 3 Indonesia is one of the telecommunications companies that has been operating in Indonesia for a long time. Even though it has been operating for a long time, there are still many users who have various different perceptions and opinions about the services provided by provider 3 Indonesia. The process of retrieving data on twitter uses the twitter API by crawling.

Algoritma The ones used in this study are Naïve Bayes Classifier. Naive Bayes is one of the simplest algorithms to apply Bayesian rules with several advantages, namely very efficient, requires little experimental data, is easy to implement and has relatively high accuracy[2].

In previous studies comparing the performance of the naïve bayes with the SVM method (Support Vector Machine) on Twitter sentiment analysis[3]. The results of the study show that the method of Naïve Bayes has a higher level of accuracy than the SVM on the analysis. This algorithm was chosen because it can be used with only a small amount of experimental data. The data used in this study are tweet which contains the keyword "triIndonesia" during a certain period in Twitter. After that, a preprocessing process is carried out to clean the data from punctuation marks and unimportant words. Then sentiment labeling is carried out on each tweet and model training was carried out Naïve Bayes Classifier.

As a telecommunications company operating in Indonesia, Tri Indonesia also complies with various regulations and

policies from the government, such as rules regarding the 4G network and restrictions on telecommunication service tariffs. The company also has stiff competition with other mobile operators in Indonesia, such as Telkomsel, XL Axiata, and Indosat Ooredoo. In a survey in 2022, it was found that Tri users in Indonesia ranked 4th most in Indonesia with a percentage of 14.8%[4]. frequency of the appearance of words or terms in documents. In TF-IDF weighting, words that often appear in a document but rarely appear in the entire document collection will have a higher weight, while words that rarely appear in a document but often appear in the entire document collection will have a lower weight[5].

2. THEORETICAL FOUNDATION

2.1 Literature Study

Previous studies on sentiment analysis, namely:

In a study on the sentiment analysis of movie opinions of twitter users in Indonesia[6]. In this study, the use of the algorithm Naïve Bayes to analyze the sentiment of movie opinions on Twitter by categorizing sentiment into positive, negative, and neutral sentiments with the conclusion that the majority of opinion sentiment about movies taken from Twitter is positive.

In his research, A. Rahman Isnain, A. Indra Sakti, D. Alita[7] about analyzing public sentiment about the lockdown policy implemented in Jakarta using an algorithm SVM. From this study, it was found that the lockdown implemented in Jakarta had more negative sentiment than positive sentiment.

In his research, A. Adhi Putra[8] entitled Sentiment analysis on user reviews of the Bibit and Bareksa application using the KNN to analyze sentiment on app user reviews Seed and

Bareksa. In this study, it was produced that the algorithm KNN It is quite effective in analyzing user reviews, while the results of the analysis show that the majority of reviews are positive. The researcher also hopes that the results of the research can provide benefits for the company[8].

In the journal Sentiment Analysis of Hate Speech in the 2019 Presidential Election Using Naïve Bayes' Algorithm[9] discussing the sentiment of hate speech related to the 2019 presidential election in Indonesia with an algorithm Naïve Bayes Classifier. Use of algorithms Naïve Bayes It is quite effective in analyzing sentiment in hate speech and can distinguish speech that contains positive, negative, or neutral sentiments. The result is that the majority of hate speech related to the 2019 presidential election in Indonesia is negative.

In his research, N. Astari, Dewa Gede Hendra Diviyana, Gede Indrawan[10] Regarding the sentiment analysis of twitter documents regarding the impact of the coronavirus using the Naïve Bayes Classifier method. This research uses an algorithm naïve bayes classifier In analyzing the sentiment of documents twitter by distinguishing positive, negative or neutral sentiments. The results of the analysis show that the majority of twitter documents related to the impact of the coronavirus are negative.

In his research, M. Umi Rofiqoh1, Rizal Setya Perdana, M. Ali Fauzi3[11] on the journal Sentiment Analysis of User Satisfaction Levels of Indonesian Mobile Telecommunication Service Providers on Twitter with Support Vector Machine Method and Lexicon Based Features Discuss the use of the SVM and lexicon feature to analyze sentiment on the level of satisfaction of Indonesian mobile telecommunications service users on twitter. Method SVM And lexicon-based features are quite effective in analyzing sentiment on user tweets and can distinguish positive, negative, or neutral sentiment. The result is that most user tweets are neutral, but some also contain positive or negative sentiments

2.2 Provider 3 Indonesia

As a telecommunications company operating in Indonesia, Tri also complies with various regulations and policies from the government, such as rules regarding 4G networks and restrictions on telecommunication service tariffs. The company also has stiff competition with other mobile operators in Indonesia, such as Telkomsel, XL Axiata, and Indosat Ooredoo. In a survey in 2022, it was found that Tri users in Indonesia ranked 4th most in Indonesia with a percentage of 14.8%[4]

2.2 Twitter API

API (Application Programming Interface) is Crawles connected tools provided by Twitter. The API in accessing Twitter data can use two access methods, namely the REST API and the Streaming API. To access the Twitter API can only be through an authentication request. Each request sent must be made by an authorized Twitter user himself and his access is limited to a certain number called rate limit. The limits used are adjusted to the level of the application and its use[12].

2.3 Text Mining

Text mining is the process of extracting information or knowledge from text or documents automatically using techniques from fields such as data mining, machine learning, dan NLP[13]. The purpose of text mining is to identify patterns and information hidden in the text or document being analyzed.

2.4 Text Preprocessing

Text preprocessing is a technique to prepare text data before analysis. The goal is to eliminate irrelevant information and clarify the meaning of the text, so that data that was initially unstructured becomes structured for processing. Some of the processes in the text preprocessing stage include data cleaning, case folding, tokenizing, stopword removal, normalization, and stemming.

2.5 TF-IDF Weighting

TF-IDF Weighting (Term Frequency-Inverse Document Frequency) It is a technique in text mining to give value weight to the words contained in a document. Needed Term Frequency (TF) is the number of words or terms that appear in a document, and Inverse Document Frequency (IDF) is the frequency of the appearance of words or terms in documents. In TF-IDF weighting, words that often appear in a document but rarely appear in the entire document collection will have a higher weight, while words that rarely appear in a document but often appear in the entire document collection will have a lower weight[5].

2.6 SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) is one of the oversampling techniques used to deal with class imbalances in datasets. SMOTE works by identifying minority sample data by creating a new synthetic sample based on its closest neighbors so that the data is balanced with the majority class data. The use of the smote method allows the existence of the same training data, because in minority classes it is duplicated.

2.7 Naïve Bayes Classifier

Method Naïve Bayes Classifier is a classification method that uses probability and statistical methods used and discovered by British scientist Thomas Bayes. Naive Bayes Classifier is a probabilistic classification model based on Bayesian probability theory by estimating the probability of an instance belonging to a class based on the set of attributes (features) it has[14].

3. RESEARCH METHODOLOGY

The Classification System carried out for sentiment analysis has a design of how the flow of this system will run. An overview of the system to be created is as shown in figure 1.

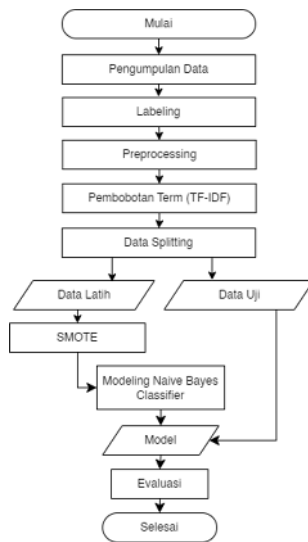


Figure 1. System Design

3.1 Research Flow

The first stage is the collection of tweet data using APIs. Data collection was carried out by crawling with the twitter API with the keyword "triindonesia". Then it was followed by manual labeling and the deletion of tweets that were not needed. Followed by the process of preprocessing, word weighting, and data sharing. Furthermore, the data is balanced using SMOTE and then the data is modeled. After that, the data is evaluated to find the best modeling.

1.2 Data Collection Process

The data collection process is carried out by crawling using python, tweepy and pandas libraries on google colab. Crawling was carried out from May 5 to May 24, 2023 and the final dataset results were obtained from 880 data. Furthermore, the data is labeled manually with 3 classifications of positive, negative and neutral sentiment. The results were obtained 578 negative, 120 positive, 183 neutral. It is shown in table 1:

Table 1. Data Collection

Tweet text	Label
It's already dm but the one who replies to the bot, the service is not good 🙄🙄🙄	Negative
Hello, how do I use my local quota? Please enlighten the bestie https://t.co/wkzKkh0t2N	Neutral
Wkwkwk, we are looking for promo packages continue to yaaa from tri	Positive

3.3 Preprocessing

In this preprocessing, there are several stages that are carried out, namely, Case Folding, Tokenizing, Stopword, Normalization and Stemming.

In the case folding process, the data is converted into all small letters. Furthermore, in the tokenizing process, the url, username, emoticon (tweet special) are deleted. The tokenization process is continued, namely the separation of words based on spaces found into a single token. In the stopword removal process, unimportant words such as connecting words, time, and others are removed. Subsequently, the normalization process was carried out by changing the non-standard word, the word slank to a standard word using a normalization dictionary. In this process, punctuation is also removed. Followed by the stemming process where this process is to change the word of the compound to the basic word.

3.3 TF-IDF Weighting

The tf-idf weighting aims to provide a value to each term and this value will be used for the classification process later. This process is done automatically using a library program in python

4. RESULT AND DISCUSSION

4.1 Splitting data

In this study, 3 experiments were carried out with the division of 90% training data, 70% training data, and 50% training data.

4.2 SMOTE Implementation

Because the class data obtained was not balanced, an upsampling process was carried out to balance the class data using SMOTE SMOTE (Synthetic Minority Over-sampling Technique). The data that is balanced is data training. Then the total training data is obtained according to table 2:

Table 2. Test data

Testing	SMOTE	Total data
90% training	516	1548
70% training	404	1212
50% training	282	846

4.2 Naïve Bayes Classifier

In this process, Naïve Bayes classifier classification is carried out, in this process it requires data that has gone through the weight of word occurrence from TF-IDF, data sharing, and SMOTE to balance the data because the dataset is not balanced. Furthermore, Naïve Bayes Classifier predicts test data from training data. After classification using naïve bayes, the results are obtained in table 3;

Table 3. Test Results

Data testing	Accuracy	Precision	Recall
10%	79,77	79,77	79,77
30%	73,58	73,58	73,58
50%	70,58	70,58	70,58

From 3 experiments with different data sharing, it was found that the best performance was in testing using 10% data testing with an accuracy of 79%.

4.2 Comparison using Textblobs

In this test, the aim is to compare the results of the implementation of the naïve bayes classifier algorithm using textblobs and without using textblobs. The goal is to find out if the use of textblobs significantly increases the accuracy value. Testing was carried out using a comparison of 90% of the training data.

The number of data used was 880 data with different classification results, where data without textblobs were given manual labels while data using textblobs used automatic labeling. The following is a comparison table of positive, negative and neutral data.

Table 4. Labeling results

Labeling Techniques	Positive	Negative	Neutral
Labeling manual	120	578	183
With Textblob	185	259	429

Furthermore, the data was tested using the naïve Bayes classifier algorithm and the results were obtained in the following table 5:

Table 5. Test results

Testing	Accuracy	Precision	Recall
No Textblob	65,82	65,82	65,82
With Textblob	63,64	63,64	63,64

4.2 Comparison without SMOTE

Tests are carried out before using the smote and after using the smote. By comparing the values of accuracy, recall, precision. This test is to prove whether SMOTE is proven to improve the performance of the naïve bayes classification. Testing was carried out using 90% of training data from 880 datasets with the following data composition:

Table 6. Labeling Results

Labeling techniques	Positive	Negative	Neutral
Labeling manual	120	578	183
With SMOTE	516	516	516

It is found in table 7 that the difference in data composition without SMOTE and with SMOTE is in the majority of data, without SMOTE the majority of data is negative. Meanwhile, in the use of SMOTE, data is equated with majority data. Next, the classification process was carried out using the naïve Bayes classifier algorithm and the results were obtained:

Table 7. Test results

Testing	Accuracy	Precision	Recall
Without SMOTE	65,82	65,82	65,82
With SMOTE	79	79	79

In this process, the highest accuracy value was obtained after using SMOTE to balance the data by 79%. Therefore, SMOTE has proven to be able to improve the performance of the naïve bayes classifier classification.

5. CONCLUSIONS

After conducting a sentiment analysis of the 3 Indonesia provider service on social media twitter with 3 dataset tests, conclusions can be drawn.

1. The majority of provider 3 indonesia's user sentiment on twitter is negative sentiment with a percentage of 65.6% of the total dataset in the data comparison experiment of 90% training data and 10% data testing.
2. From the tests that have been carried out, there are 3 tests with different data comparisons. The highest performance results were obtained in testing 90% of the training data with an accuracy score of 79%.

REFERENCE

- [1] S. Kemp, "Digital 2022: Indonesia — DataReportal — Global Digital Insights," *Global Digital Insights*. pp. 1–103, 2022, Accessed: Mar. 14, 2023. [Online]. Available: <https://datareportal.com/reports/digital-2022-indonesia>.
- [2] M. Ghifari *et al.*, "ANALISIS SENTIMEN TWITTER TERHADAP KENAIKAN BAHAN SENTIMENT ANALYSIS OF TWITTER ON FUEL OIL INCREASE USING," vol. 2, no. April, pp. 219–226, 2023.
- [3] M. I. Fikri, T. S. Sabrila, and Y. Azhar, "Perbandingan Metode Naïve Bayes dan Support Vector Machine pada Analisis Sentimen Twitter," *SMATIKA J.*, vol. 10, no. 02, pp. 71–76, 2020, doi: 10.32664/smatika.v10i02.455.
- [4] "5 Operator Seluler Favorit Masyarakat Indonesia Versi APJII." <https://databoks.katadata.co.id/datapublish/2022/06/13/5-operator-seluler-favorit-masyarakat-indonesia-versi-apjii> (accessed Mar. 15, 2023).
- [5] V. Amrizal, "Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Hadits Shahih Bukhari-Muslim)," *J. Tek. Inform.*, vol. 11, no. 2, pp. 149–164, 2018, doi: 10.15408/jti.v11i2.8623.
- [6] F. Ratnawati, "Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter," *INOVTEK Polbeng - Seri Inform.*, vol. 3, no. 1, p. 50, 2018, doi: 10.35314/isi.v3i1.335.

- [7] A. Rahman Isnain, A. Indra Sakti, D. Alita, and N. Satya Marga, "Sentimen Analisis Publik Terhadap Kebijakan Lockdown Pemerintah Jakarta Menggunakan Algoritma Svm," *Jdmsi*, vol. 2, no. 1, pp. 31–37, 2021, [Online]. Available: <https://t.co/NfhnfMjtXw>.
- [8] A. D. Adhi Putra, "Analisis Sentimen pada Ulasan pengguna Aplikasi Bibit Dan Bareksa dengan Algoritma KNN," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 8, no. 2, pp. 636–646, 2021, doi: 10.35957/jatisi.v8i2.962.
- [9] M. Chalida and M. D. R. Wahyudi, "Analisis Sentimen Ujaran Kebencian Pemilihan Presiden 2019 Menggunakan Algoritma Naïve Bayes (Studi Kasus: Tweet #Pilpres2019 Di Kota Jakarta, Bandung, Semarang, Surabaya Dan Yogyakarta)," *Jnanaloka (Jurnal Open Access Yayasan Lentera Dua Indones.)*, no. 2001, pp. 5–10, 2019.
- [10] N. M. A. J. Astari, Dewa Gede Hendra Divayana, and Gede Indrawan, "Analisis Sentimen Dokumen Twitter Mengenai Dampak Virus Corona Menggunakan Metode Naïve Bayes Classifier," *J. Sist. dan Inform.*, vol. 15, no. 1, pp. 27–29, 2020, doi: 10.30864/jsi.v15i1.332.
- [11] M. A. F. Umi Rofiqoh¹, Rizal Setya Perdana² and Program, "Analisis Sentimen Tingkat Kepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonesia Pada Twitter dengan Metode Support Vector Machine dan Lexicon Based Features Twitter event detection View project Human Detection and Tracking View project," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1(12), no. October, pp. 1725–1732, 2017, [Online]. Available: <https://www.researchgate.net/publication/320234928>.
- [12] E. S. Negara, R. Andryani, and P. H. Saksono, "Analisis Data Twitter: Ekstraksi dan Analisis Data Geospasial," *J. INKOM*, vol. 10, no. 1, p. 27, 2016, doi: 10.14203/j.inkom.433.
- [13] A. Rossi, T. Lestari, R. Setya Perdana, and M. A. Fauzi, "Analisis Sentimen Tentang Opini Pilkada DKI 2017 Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naïve Bayes dan Pembobotan Emoji," *Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 12, pp. 1718–1724, 2017, [Online]. Available: <http://j-ptiik.ub.ac.id>.
- [14] I. H. Witten, M. A. Hall, and E. Frank, "Data Mining: Practical Machine Learning Third Edition," (*Morgan Kaufmann Ser. Data Manag. Syst. Morgan Kaufmann*, vol. 104, no. June, p. 113, 2005.