Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

June 2026. Vol. 5. No. 3; e-ISSN: 2808-4519

Implementation of Naïve Bayes Algorithm for Sentiment Classification of Public Youtube Opinions Related to Nickel Mining Issues in Raja Ampat

Wanda Arfilla Daulay[1]*, Relita Buaton[2], Kristina Annatasia Br Sitepu[3]

[1,2,3]*Program of Study Information System, STMIK Kaputama*
wandaarfilla988@gmail.com[1*], bbcbuaton@gmail.com[2], kannatasia88@gmail.com[3]

Abstract

Indonesia currently holds the world's largest nickel reserves. However, extractive activities in the Raja Ampat conservation area pose significant ecological threats and trigger social polarization on social media. The massive volume of opinion data creates challenges for policymakers in mapping public perception quickly and objectively. This study aims to classify public sentiment regarding nickel mining activities in Raja Ampat using the Naïve Bayes algorithm with TF-IDF feature weighting. The methodology employed in this research is CRISP-DM, which consists of Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment stages. The dataset consists of 10,903 YouTube comments collected from the Ferry Irwandi and Kompas.com channels. The results indicate that negative sentiment dominates public discourse at 46.4%, followed by neutral sentiment at 31%, and positive sentiment at 22.6%. The classification model achieved an accuracy rate of 71.59%. Furthermore, a Streamlit-based visualization dashboard was developed to assist stakeholders in monitoring public opinion trends systematically.

Keywords: *Naïve Bayes; Nickel Mining; Raja Ampat; Sentiment Analysis; YouTube*

1. Introduction

Indonesia currently holds a vital position on the global industrial map as the country with the largest nickel reserves in the world. This commodity is no longer just a raw natural resource, but has transformed into a strategic backbone for the modern industrial ecosystem, especially in supporting the global trend of the electric vehicle revolution and battery technology. This shift in function makes nickel a high-value economic commodity that encourages massive industrial activity through the direction of national downstream policies [1].

However, the expansion of nickel mining activities has created environmental concerns, particularly in Raja Ampat, Southwest Papua, which is internationally recognized as one of the world's richest marine biodiversity regions. Mining operations in this conservation area potentially cause deforestation, coastal sedimentation, and coral reef degradation that threaten ecosystem sustainability and local community livelihoods [2].

The environmental and social impacts of mining activities have triggered intense public discussions on social media platforms. YouTube has become one of the primary digital platforms where users freely express opinions, criticisms, and support regarding public issues. According to Modami et al., YouTube serves as an important medium for digital interaction and public opinion exchange in Indonesia due to its massive number of active users [3].

To overcome these challenges, sentiment analysis can be implemented as an automated approach to classify public opinions into positive, negative, and neutral categories. Sentiment analysis is part of Natural Language Processing (NLP), which enables computers to understand and process textual information efficiently [2]. In this study, the Naïve Bayes algorithm was selected because of its simplicity, computational efficiency, and strong performance in text classification involving high-dimensional textual data [4].

This research also applies the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology to ensure systematic and structured data processing stages, including Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment [5]. Therefore, this study is expected to provide valuable insights regarding public sentiment toward nickel mining activities in Raja Ampat and support environmental conservation policies.

2. Research Methodology

2.1. Text Mining

Text mining is the process of extracting useful information and knowledge from unstructured text data. Text mining can efficiently analyze comprehensive information about policy discourse by examining various issues discussed in reality to determine related sub-policy keywords [6].

2.2. Sentimen Analysis

Sentiment analysis is classifying each pole of text on various internet and social media sources in the form of documents or sentences, then determining whether the word belongs to the positive, neutral, or negative category[7]. Sentiment analysis is also capable of showing emotions such as sadness, joy, or anger contained in the text data.

2.3. Naïve Bayes

Naïve Bayes is a probability-based classification algorithm that works based on Bayes Theorem to predict the class of a data by calculating the probability of occurrence based on the features it has [1]. This algorithm is called naïve because it assumes that each feature is independent of each other, a simple approach that while not always fulfilled under real conditions, proves to be very effective and efficient. This method is particularly reliable in the processing of text data such as sentiment analysis, document classification, and spam filtering, where the assumption of independence actually simplifies the calculation process on data with a large number of features without significantly reducing performance. Mathematically, Naïve Bayes is based on Bayes' Theorem, which is formulated as follows:

$$P(c|d)=\frac{P(d|c) \times P(c)}{P(d)} \quad \ldots\ldots(1)$$

2.4. Cross Industry Standard Process For Data Mining

CRISP-DM (Cross-Industry Standard Process for Data Mining) is a process model that is widely used in data science due to its technology-independent nature and adaptability in various industries. This model defines key steps to ensure the project runs systematically [5]

2.5. Python

Python is a very popular high-level programming language, known for its simple and easy-to-understand syntax [8]. First introduced by Guido van Rossum in 1991, Python has grown to become one of the most widely used programming languages in the world, especially in the fields of data science, web development, and automation. Python supports a wide range of programming paradigms, including procedural, object-oriented and functional programming.

3. Results and Discussion

This research method uses the CRISP-DM (Cross Industry Standard Process for Data Mining) framework as a systematic and structured guide. The research stages include data scrapping, crawling, data preprocessing, labeling data, split data, and Naive Bayes model evaluation. Figure 1 shows the research stages, which include:

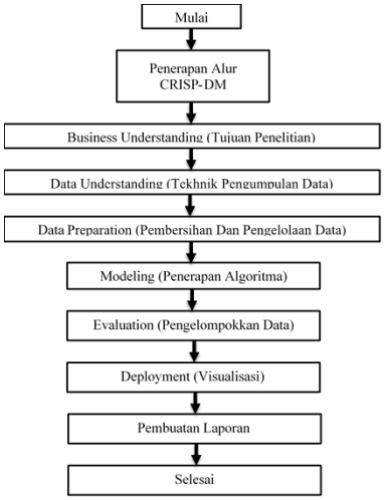

Fig 1: Research Stages

3.1. Scraping Data

Data scraping is a method of collecting data automatically that aims to retrieve specific information from a website or digital platform. The data obtained is then processed and stored in a structured form so that it is ready to be used for further analysis processes [9].

The data used for this research comes from comments on Ferry Irwandi YouTube channel. Python was used to collect 10.753 comments. The data is then processed beforehand, before further analysis

Table 1: Scraping

Author	Comment_Text	Published_at
@imassuryati8471	Sampai sekarang sudah adat yg penting komisi ÃƒÂ°Ã,ÂŸÃ,Â˜Ã,Â,ÃƒÂ°Ã,ÂŸÃ,Â˜Ã,Â,ÃƒÂ°Ã,ÂŸÃ,Â˜Ã,Â,ÃƒÂ°Ã,ÂŸÃ,Â˜Ã,Â,ÃƒÂ°Ã,ÂŸÃ,ÂŽÃ,Â‰ÃƒÂ°Ã,ÂŸÃ,ÂŽÃ,Â‰ÃƒÂ°Ã,ÂŸÃ,ÂŽÃ,Â‰	2026-01-03
@Adangdarusalamchannel	Harga recovery lebih mahal,, dari pada ke untungan tambang	2025-12-03
@GonoAldi	Oo wong solo dari dahulu kala banyak orang solo di pemerintahan ,/ politikus eh jadi tikus beneran	2025-09-08

3.2. Pre-Processing Data

This stage aims to clean the text data of elements that can interfere with the performance of the Naïve Bayes algorithm

1. Case Folding
   Case folding is the stage of converting all letters in text into lowercase letters. This is done to ensure data consistency, so that the same words but with different capital letters such as "Mine" and "mine" will be considered as one identical word by the system. The case folding results are shown in Table 2.

Table 2: Case Folding

No	Before	After
1	Sampai sekarang sudah adat yg penting komisi ÃƒÂ°Ã,ÂŸÃ,Â˜Ã,Â,ÃƒÂ°Ã,ÂŸÃ,Â˜Ã,Â,ÃƒÂ°Ã,ÂŸÃ,Â˜Ã,Â,ÃƒÂ°Ã,ÂŸÃ,Â˜Ã,Â,ÃƒÂ°Ã,ÂŸÃ,ÂŽÃ,Â‰ÃƒÂ°Ã,ÂŸÃ,ÂŽÃ,Â‰ÃƒÂ°Ã,ÂŸÃ,ÂŽÃ,Â‰	sampai sekarang sudah adat yg penting komisi Ã£Â°Ã¢ÂŸÃ¢Â˜Ã¢Â,Ã£Â°Ã¢ÂŸÃ¢Â˜Ã¢Â,Ã£Â°Ã¢ÂŸÃ¢Â˜Ã¢Â,Ã£Â°Ã¢ÂŸÃ¢ÂŽÃ¢Â‰Ã£Â°Ã¢ÂŸÃ¢ÂŽÃ¢Â‰Ã£Â°Ã¢ÂŸÃ¢ÂŽÃ¢Â‰
2	Harga recovery lebih mahal,, dari pada ke untungan tambang	harga recovery lebih mahal,, dari pada ke untungan tambang
3	Oo wong solo dari dahulu kala banyak orang solo di pemerintahan ,/ politikus eh jadi tikus beneran	oo wong solo dari dahulu kala banyak orang solo di pemerintahan ,/ politikus eh jadi tikus beneran

2. Cleaning Data
   This stage involves removing elements that have no meaning in the sentiment analysis, such as links (URLs), mentions, punctuation, symbols, and numbers that appear in YouTube comments about the mining issue in Raja Ampat. The results of data cleansing are in table 3.

Table 3: Cleaning Data

No	Before	After
1	sampai sekarang sudah adat yg penting komisi Ã£Â°Ã¢ÂŸÃ¢Â˜Ã¢Â,Ã£Â°Ã¢ÂŸÃ¢Â˜Ã¢Â,Ã£Â°Ã¢ÂŸÃ¢Â˜Ã¢Â,Ã£Â°Ã¢ÂŸÃ¢ÂŽÃ¢Â‰Ã£Â°Ã¢ÂŸÃ¢ÂŽÃ¢Â‰Ã£Â°Ã¢ÂŸÃ¢ÂŽÃ¢Â‰	sampai sekarang sudah adat yg penting komisi
2	harga recovery lebih mahal,, dari pada ke untungan tambang	harga recovery lebih mahal dari pada ke untungan tambang
3	oo wong solo dari dahulu kala banyak orang solo di pemerintahan ,/ politikus eh jadi tikus beneran	oo wong solo dari dahulu kala banyak orang solo di pemerintahan politikus eh jadi tikus beneran

3. Tokenizing
   Tokenizing is the process of cutting or breaking sentences into single word units (tokens). Through this process, each community's comment will be broken down into a collection of words whose frequency of occurrence will later be calculated by the classification model. The results of the tokenizing are shown in Table 4.

Table 4: Tokenizing

No	Before	After
1	sampai sekarang sudah adat yg penting komisi	['sampai', 'sekarang', 'sudah', 'adat', 'yg', 'penting', 'komisi']
2	harga recovery lebih mahal dari pada ke untungan tambang	['harga', 'recovery', 'lebih', 'mahal', 'dari', 'pada', 'ke', 'untungan', 'tambang']
3	oo wong solo dari dahulu kala banyak orang solo di pemerintahan politikus eh jadi tikus beneran	['oo', 'wong', 'solo', 'dari', 'dahulu', 'kala', 'banyak', 'orang', 'solo', 'di', 'pemerintahan', 'politikus', 'eh', 'jadi', 'tikus', 'beneran']

4. Word Normalization
The normalization process is carried out to improve the quality of text data by changing abbreviations and non-standard words that often appear on YouTube into standard words, thereby improving the accuracy of sentiment analysis.The results for Word Normalization are shown in table 5.

Table 5: Word Normalization

No	Before	After
1	['sampai', 'sekarang', 'sudah', 'adat', 'yg', 'penting', 'komisi']	['sampai', 'sekarang', 'sudah', 'adat', 'yang', 'penting', 'komisi']
2	['harga', 'recovery', 'lebih', 'mahal', 'dari', 'pada', 'ke', 'untungan', 'tambang']	['harga', 'recovery', 'lebih', 'mahal', 'dari', 'pada', 'ke', 'untungan', 'tambang']
3	['oo', 'wong', 'solo', 'dari', 'dahulu', 'kala', 'banyak', 'orang', 'solo', 'di', 'pemerintahan', 'politikus', 'eh', 'jadi', 'tikus', 'beneran']	['oo', 'wong', 'solo', 'dari', 'dahulu', 'kala', 'banyak', 'orang', 'solo', 'di', 'pemerintahan', 'politikus', 'eh', 'jadi', 'tikus', 'benar']

5. Stopword Removal
Stopword removal aims to eliminate common words that appear frequently but do not have a significant effect on sentiment, such as the words "which", "in", "and", "for", or "to". The removal of these words is done so that the algorithm can focus more on words that contain emotional value or specific opinions about mining policy. The results for Stopword Removal are shown in table 6.

Table 6: Stopword Removal

No	Before	After
1	['sampai', 'sekarang', 'sudah', 'adat', 'yang', 'penting', 'komisi']	['adat', 'komisi']
2	['harga', 'recovery', 'lebih', 'mahal', 'dari', 'pada', 'ke', 'untungan', 'tambang']	['harga', 'recovery', 'mahal', 'untungan', 'tambang']
3	['oo', 'wong', 'solo', 'dari', 'dahulu', 'kala', 'banyak', 'orang', 'solo', 'di', 'pemerintahan', 'politikus', 'eh', 'jadi', 'tikus', 'benar']	['oo', 'solo', 'solo', 'pemerintahan', 'politikus', 'tikus']

6. Stemming
Stemming is the process of mapping and converting various forms of words into one root *word*. In this study, *the stemming process* was carried out using the help of a library specifically for the Indonesian language (such as Sastrawi). The main goal of this stage is to eliminate suffix in the form of prefixes, inserts, suffixes, or combinations of the three, so that the frequency of occurrence of a root word can be calculated more accurately by the *Naïve Bayes algorithm.* The stemming results are shown in Table 7.

Table 7: Stemming

No	Before	After
1	['adat', 'komisi']	adat komisi
2	['harga', 'recovery', 'mahal', 'untungan', 'tambang']	harga recovery mahal untung tambang
3	['oo', 'solo', 'solo', 'pemerintahan', 'politikus', 'tikus']	oo solo solo perintah politikus tikus

3.3. Data Labeling

Labeling is the process of marking or categorizing each text data, such as Positive, Negative, or Neutral. This process is very important in the development of machine learning models using the Naïve Bayes algorithm, because the algorithm requires labeled training data to learn the word patterns that represent each sentiment. The determination of labels in this study was carried out based on word weighting using a sentiment dictionary (lexicon based). If the total weight of words in a sentence is greater than zero (>0) then it is categorized as Positive, if it is less than zero (<0) as Negative, and if it is equal to zero (0) then it is categorized as Neutral. Here is an example of the weight calculation in comment number 3 in the labeling process. The results of the data labeling are shown in Table 8.

Table 8: Data Labeling

Comments	Sentiment
['adat', 'komisi']	Positive
['harga', 'recovery', 'mahal', 'untungan', 'tambang']	Positive
['oo', 'solo', 'solo', 'pemerintahan', 'politikus', 'tikus']	Positive

3.4. Splitting Data

After all the data of the public sentiment review text through the preprocessing and word weighting stages using TF-IDF, the next stage in this study is data splitting. Data sharing is a crucial step in building a machine learning model to evaluate performance and prevent overfitting or underfitting in the classification model. In this study, opinion data regarding the existence of nickel mines in Raja Ampat was divided by a ratio of 70:30, where 70% of the data was used as system learning material (Latihan Data) and the remaining 30% was used as testing material (Test Data). This division is important to obtain an objective assessment of the model's performance through the Accuracy, Precision, Recall, and F1-Score scores. The results are shown in Table 9.

Table 9: Splitting Data

Rasio Perbandingan	Data Training	Data Testing
70:30	7.527	3.226

3.5. Classification Results

TF-IDF menghitung bobot akhir dari setiap kata untuk setiap komentar (tweet/opini). Hasil dari perhitungan ini mengubah teks yang bersifat kualitatif menjadi vektor angka (kuantitatif) agar dapat diproses secara matematis oleh sistem. Nilai TF-IDF didapatkan dengan mengalikan nilai Term Frequency (TF) dengan Inverse Document Frequency (IDF). The results are shown in Table 10:

1. TF-IDF Extraction

$$TF - IDF(t) = TF(t) \times IDF(t) \quad (2)$$

Table 10: TF-IDF

Kata (Term)	TF	IDF	Perhitungan (TF × IDF)	Hasil TF-IDF
harga	1,0	1,903	1,0 × 1,903	1,903
recovery	1,0	1,903	1,0 × 1,903	1,903
mahal	1,0	1,903	1,0 × 1,903	1,903
untung	1,0	1,903	1,0 × 1,903	1,903
tambang	1,0	1,903	1,0 × 1,903	1,903

2. Modelling

At this stage, the data that has gone through the TF-IDF preprocessing and weighting process will be processed using the Multinomial Naïve Bayes algorithm. This algorithm works based on probability theory to determine which class (Positive, Negative, or Neutral) an opinion tends to enter based on the words contained in it. The training data is used to calculate the initial probability (Prior) and the probability of the word to the class (Likelihood). Here is a table 11 of features from the training data:

Table 11: Modeling (Bobot TF-IDF per Category)

No	x1 (Total Bobot Negatif)	x2 (Total Bobot Positif)	x3 (Total Bobot Netral)	Sentimen (Label)
T1	0	3,630	0	Positif (+1)
T2	0	9,515	0	Positif (+1)
T3	0	6,902	0	Positif (+1)
T4	0	9,515	0	Positif (+1)
T5	18,428	0	0	Negatif (-1)
T6	9,155	0	0	Negatif (-1)
T7	0	26,908	0	Positif (+1)
T8	0	0	3,105	Netral (0)
T9	0	5,709	0	Positif (+1)
T10	0	0	5,709	Netral (0)

3. Calculation of Bayes Naive Probability

Naïve Bayes works on the basis of the approach of probability theory. This model calculates Prior Probability (the initial chance of a class appearing) and Likelihood (the chance of a feature/word appearing in the class) The results are shown in Table 12

Table 12: Total Bobot Category (W)

Kategori	Simbol	Perhitungan (Penjumlahan Bobot Latih)	Total Bobot TF-IDF
Positif	$\sum x_2$	3,630 + 9,515 + 6,902 + 9,515 + 26,908 + 5,709	62,179
Negatif	$\sum x_1$	18,428 + 9,155	27,583
Netral	$\sum x_3$	3,105 + 5,709	8,814

The test data was then calculated for its probability value for each class using the Naive Bayes algorithm. The prediction results are determined based on the highest probability value (ArgMax) obtained from the posterior probability calculation, which is the result of multiplication between the prior probability value and likelihood in each class. The class with the greatest probability value is selected as the result of the final classification. Based on the results of the tests on five samples of test data, the Naive Bayes algorithm managed to correctly classify the entire document. This is indicated by the conformity between the actual class label and the system prediction results on each test data. The highest probability value (ArgMax) generated always points to the category that corresponds to the actual label, so that all predictions made by the system are declared correct. The results are shown in Table 13

Table 13: Prediction Calculation Results (Posterior)

No	Nilai Negatif (P×x1)	Nilai Positif (P×x2)	Nilai Netral (P×x3)	Prediksi Sistem	Label Asli	Hasil
T11	0	0	1,903	Netral	Netral	Sesuai
T12	5,982	0	0	Negatif	Negatif	Sesuai
T13	13,0838	0	0	Negatif	Negatif	Sesuai
T14	2,2836	0	0	Negatif	Negatif	Sesuai
T15	7,8924	0	0	Negatif	Negatif	Sesuai

4. Classification Results

Based on the results of tests conducted on 10,903 test data, the Naïve Bayes model produced an Overall Accuracy rate of 71.59%. This achievement shows that in general, the model is able to classify seven out of ten public comments precisely into negative, neutral, or positive categories. Negative Sentiment: It has the most stable performance with a Precision value of 0.828 and a Recall of 0.725. This is supported by the most dominant number of data (support), which is 5,769 comments, so that the model has richer word pattern references to recognize the narrative of rejection or public concern. Neutral Sentiment: This category exhibits unique characteristics with high Recall (0.807) but low Precision (0.474). This figure indicates that the model is very sensitive in capturing comments that are informative or general, but there is often a "misconception" where comments from other categories (positive/negative) are classified as neutral. Positive Sentiment: Shows a fairly good balance of performance with an F1-Score of 0.719 and Precision of 0.819. This proves that when the system predicts a comment as a positive sentiment, the level of truth is very high. Overall, the Macro Average F1-Score value of 0.697 and the Weighted Average F1-Score of 0.726 show that the Naïve Bayes model is quite resilient and reliable in mapping public opinion trends.

Despite the challenges to the precision of neutral sentiment, the high precision values in the negative and positive categories ensure that the results of this analysis can be accounted for to assist stakeholders in objectively mapping people's aspirations.

Matriks Evaluasi Kinerja Algoritma

Akurasi Keseluruhan Model: **71.59%**

Deploy

	Presisi (Precision)	Sensitivitas (Recall)	F1-Score	Jumlah Data (Support)
Negatif	0.828	0.725	0.773	5769
Netral	0.474	0.807	0.597	1988
Positif	0.819	0.641	0.719	3146
accuracy	0.716	0.716	0.716	1
macro avg	0.707	0.724	0.697	10903
weighted avg	0.761	0.716	0.726	10903

Fig 1: Classification Report

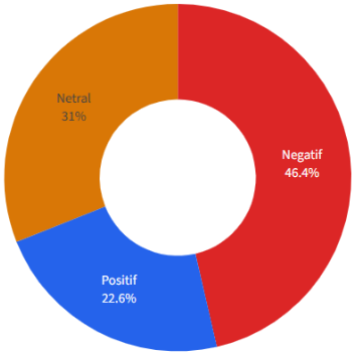

Fig 2: Sentiment Distribution Chart

3.6. Evaluation Results

At this stage, the performance of the Naive Bayes model was evaluated to measure its effectiveness in sentiment classification. Model evaluation was conducted using a confusion matrix, which provides detailed information regarding correctly classified and misclassified test data across each sentiment category. Through this evaluation, the overall model performance can be assessed using metrics such as accuracy, precision, recall, and f1-score. The evaluation results are presented as follows.

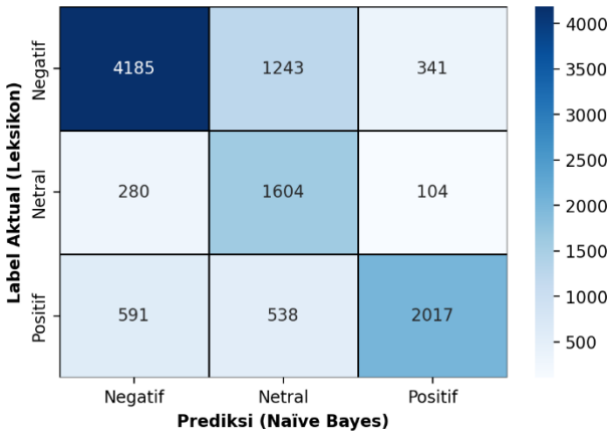

Fig 3: Confusion Matrix

3.7. Deployment

The final stage of this research was the implementation of the sentiment classification model into a Streamlit-based web application. This system was developed to provide an interactive interface that allows users to input new comment text and obtain sentiment prediction results automatically. In addition, the application also displays sentiment visualization features, such as sentiment distribution charts and evaluation results, to support easier interpretation of the classification outcomes.

Fig 4: Main Interface of Streamlit Application

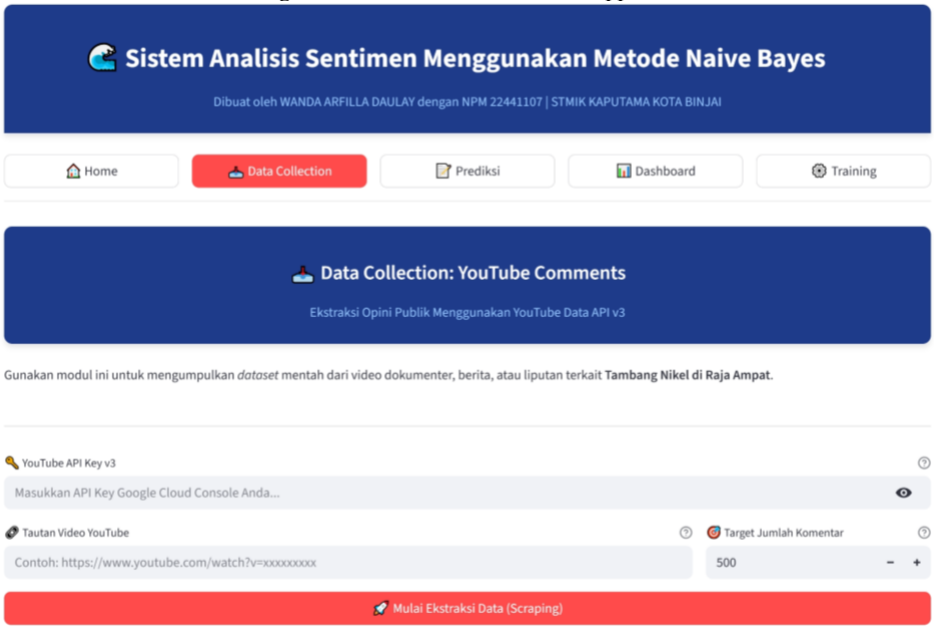

Fig 5: Data Collection

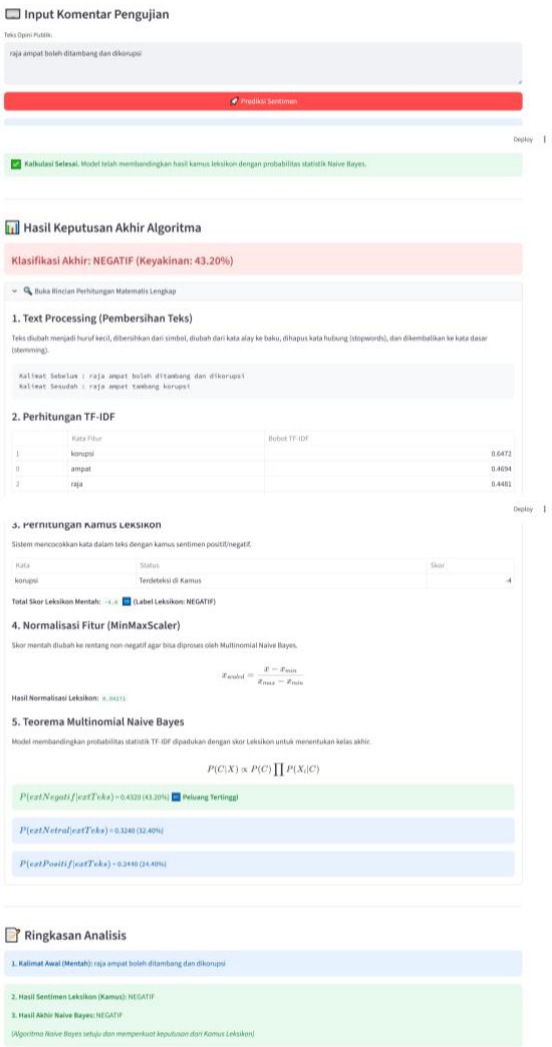

Fig 6: Sentiment Prediction Result

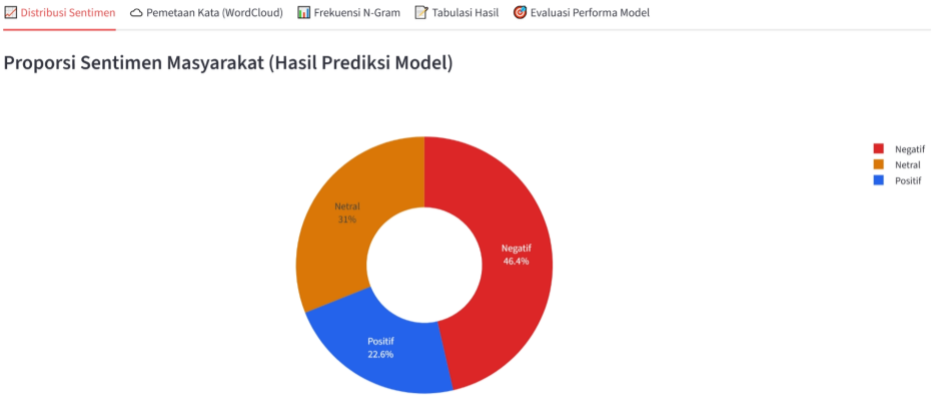

Fig 7: Sentiment Visualization Dashboard

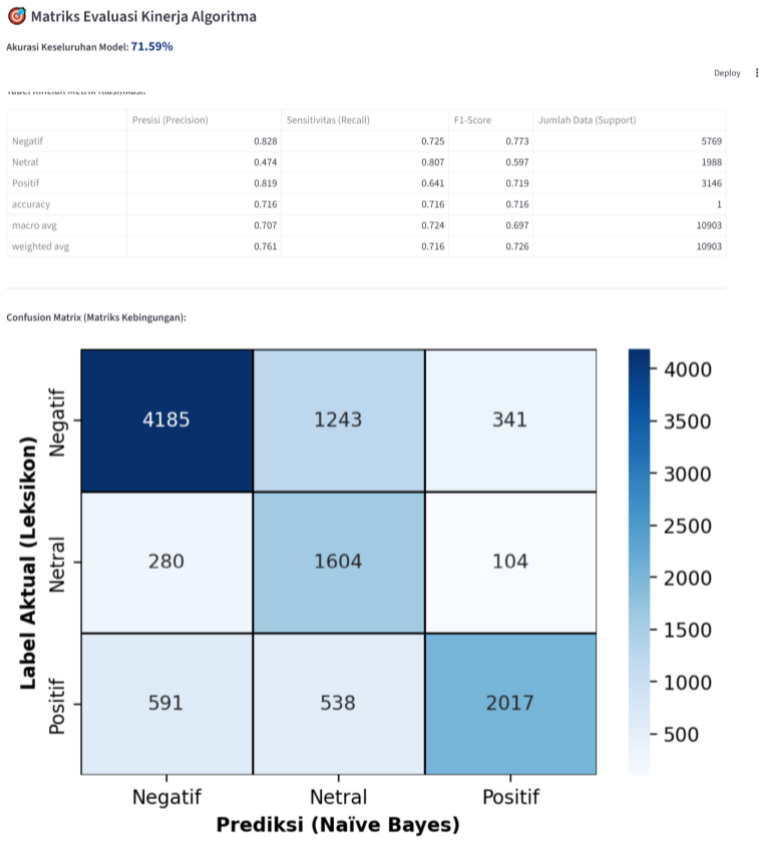

Fig 8: Model Evaluation Display

4. Conclusion

This study successfully applied the Naïve Bayes algorithm with TF-IDF weighting using the CRISP-DM framework for the sentiment analysis of YouTube comments related to the mining issue in Raja Ampat. The results of the analysis of 10,903 comments showed that negative sentiment dominated with a percentage of 46.4%, followed by neutral sentiment at 31% and positive at 22.6%, reflecting high public attention to environmental impacts and regional economic development. The developed model obtained an accuracy of 71.59% with a Macro F1-Score value of 0.697, so that it was able to provide quite good performance in mapping the overall public opinion trend.

5. Suggestion

Further research suggests applying data balancing techniques, such as SMOTE or class weight adjustment, to improve classification accuracy in neutral categories. In addition, it is necessary to compare with other methods, such as Support Vector Machine (SVM) and BERT-based Deep Learning models, in order to obtain a more optimal sentiment analysis performance on Indonesian-language social media data.

Referensi

[1] T. Agustiranti, A. Khalfani Izzati Kurdiana, B. Al Ghiffari, E. Dwi Juniar, and D. Gita Purnama, "Penerapan Naive Bayes Terhadap Sentimen Analisis Media Sosial Twitter Pengguna Kereta Cepat Jakarta-Bandung (Whoosh)," Jurnal Ilmu Komputer dan Sistem Informasi (JIKOMSI, vol. 7, no. 1, pp. 297–305, 2024.

[2] N. Suarna and W. Prihartono, "PENERAPAN NLP (NATURAL LANGUAGE PROCESSING) DALAM ANALISIS SENTIMEN PENGGUNA TELEGRAM DI PLAYSTORE," 2024.

[3] N. Modami, E. Eleazar Reva Manopo, D. Rafi Enditama, and A. Trista Ayunda, "Analisis Sentimen Komentar YouTube terhadap Rumor Peluncuran iPhone 17 Menggunakan Web Scraping dan Studi Komparatif Algoritma Klasifikasi," Journal Of Information Systems And Informatics Engineering), vol. 9, no. 2, pp. 403–411, 2025, doi: 10.35145/joisie.v9i2.5684.

[4] Nurian and B. Nurina Sari, "ANALISIS SENTIMEN ULASAN PENGGUNA APLIKASI GOOGLE PLAY MENGGUNAKAN NAÏVE BAYES," Jurnal Informatika dan Teknik Elektro Terapan, vol. 11, no. 3, pp. 2830–7062, Sep. 2023, doi: 10.23960/jitet.v11i3%20s1.3348.

[5] M. Shimaoka, R. C. Ferreira, and A. Goldman, "Leveraging XP and CRISP-DM for Agile Data Science Projects," May 2025, doi: 10.24018/ejece.2025.9.4.739.

[6] S. Do Park, "Policy Discourse Among the Chinese Public on Initiatives for Cultural and Creative Industries: Text Mining Analysis," Sage Open, vol. 12, no. 1, Mar. 2022, doi: 10.1177/21582440221079927.

[7] S. Melina Salsabila, A. Alim Murtopo, and N. Fadhilah, "Analisis Sentimen Pelanggan Tokopedia Menggunakan Metode Naïve Bayes Classifier," Jurnal Minfo Polgan, vol. 02, pp. 30–35, Sep. 2022, [Online]. Available: www.tokopedia.com

[8] Pannadhitthana Candra, "Analisis Data Menggunakan Python: Memperkenalkan Pandas dan NumPy," vol. 3, no. 1, pp. 11–16, 2025.

[9] H. Rachman, R. Megasari, and E. P. Nugroho, "Implementasi Penerapan Metode Scraping pada Pembuatan Curiculum Vitae," 2021. [Online]. Available: https://ejournal.upi.edu/index.php/JATIKOM