Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th October 2025. Vol. 5. No. 1; e-ISSN: 2808-4519

Sentiment Analysis of Students on Campus Facilities and Infrastructure Using the Naïve Bayes Classifier Method (Case Study STMIK Kaputama)

Rizky Fajar Sitepu[1]*, Relita Buaton[2], Muammar Khadapi [3]

[1,2,3]*STMIK Kaputama*

rizkyfajarsitepu@gmail.com[1*], bbcbuaton@gmail.com[2], khdafi5@gmail.com[3]

Abstract

Campus facilities and infrastructure play an important role in supporting the quality of learning. STMIK Kaputama faces challenges in maintaining the quality of its facilities as the number of students increases. This study applies sentiment analysis to student comments regarding classrooms, laboratories, libraries, restrooms, parking, and internet access. The method used is the Naïve Bayes Classifier with TF-IDF weighting and text preprocessing, following the CRISP-DM framework. The results show an accuracy of 73%, with the best performance in the positive class with precision 0.72; recall 0.97; F1-score 0.82, while the negative class with precision 0.79; recall 0.38; F1-score 0.51 and the neutral class was not detected. These findings indicate that the model tends to be dominant in positive sentiment but is still weak in distinguishing between negative and neutral comments.

Keywords: *Sentiment Analysis*, Naïve Bayes Classifier, *TF-IDF, CRISP-DM*, *Campus Facilities and Infrastructure*

1. Introduction

Facilities and infrastructure are important elements in supporting the quality of higher education. Law No. 20 of 2003 of the Republic of Indonesia requires every institution to provide adequate facilities to support student development[1]. Education produces high-quality human resources, and higher education institutions are required to provide adequate facilities to enhance learning capabilities and create an engaging academic environment[2]. As the number of students increases and academic needs grow, evaluating student perceptions becomes a crucial foundation for sustainable development.

At STMIK Kaputama, facilities such as classrooms, laboratories, libraries, toilets, parking lots, and the internet play a vital role, but students still highlight challenges such as broken chairs and air conditioners, limited collections, laboratory disruptions, toilet cleanliness, narrow parking areas, and unstable internet connections. This situation underscores the need for data-driven evaluation, including sentiment analysis as part of Natural Language Processing (NLP), which can identify positive, negative, or neutral opinions from student comments[3].

This study aims to classify student comments regarding STMIK Kaputama campus facilities using TF-IDF and the Naïve Bayes algorithm. The research findings are expected to provide an objective overview that can serve as a basis for decision-making in the continuous improvement of infrastructure.

2. Theoretical Foundation

2.1. Sentiment Analysis

Sentiment analysis is the process of automatically understanding and processing textual data to obtain information about opinions on a particular subject or object. This analysis is carried out with the aim of detecting and grouping sentiments into positive, negative, or neutral categories using Text Mining-based classification techniques[4].

2.2. Natural Langue Processing (NLP)

Natural Language Processing (NLP), a subfield of artificial intelligence, is dedicated to exploring the complex relationship between computers and human language. Its primary goal lies in understanding, examining, and generating textual information in a manner that mimics human capabilities[5].

2.3. Text Mining

Text mining is the process of discovering useful knowledge and patterns from massive and unstructured text data[6]. This method is an innovative and efficient solution for extracting hidden information from very large and complex volumes of text[7].

2.4. Text PreProcessing

Text preprocessing is the process of preparing text data to clean and standardize it in order to improve the accuracy of analysis[8].

Main Steps:

a) Cleaning: Removing distracting elements such as URLs, hashtags, and numbers.
b) Case Folding: Converting all letters to lowercase.
c) Tokenizing: Breaking sentences into word fragments (tokens).
d) Stopword Removal: Removing common, meaningless words (e.g., "the," "in," "from").
e) Normalization: Correcting non-standard words (slang, abbreviations, typos).
f) Stemming: Restoring words to their root form (e.g., "cooking" → "cook").

2.5. Classification

Text classification is the process of labeling text data into specific categories based on patterns. In the context of this study, this method was used to group students' sentiments toward campus facilities into positive, negative, or neutral categories. This process requires dividing the data into two types: training data (labeled documents for training the model) and test data (new documents for evaluating the model's performance).

2.6. Term Frequency – Inverse Document Frequency (TF-IDF)

Term Frequency – Inverse Document Frequency (TF-IDF) is a word weighting technique that calculates the frequency value of a word and the number of words that appear in the entire collection of text documents. Term Frequency (TF) refers to how often a word appears in a particular document. The higher the frequency of occurrence of the word, the greater the TF value obtained. Inverse Document Frequency (IDF) measures the number of documents containing a word compared to the total number of documents in the dataset. The rarer the word appears in various documents, the higher the IDF value[9].

2.7. *Python*

Python is a high-level, open-source programming language that is often used by programmers because of its simple and easy-to-understand syntax. These characteristics make Python one of the easiest languages to use, both for beginners and researchers. Writing code in Python follows clear and structured rules to minimize errors during execution. Additionally, Python is known as a versatile language and is widely used in various fields such as data processing, machine learning development, and text analysis, including sentiment analysis based on opinions[10].

2.8. Google Colab

Google Colab is a web-based Integrated Development Environment (IDE) from Google that uses Jupyter Notebook with the .ipynb extension. This platform supports various Python libraries such as Keras, TensorFlow, NumPy, Pandas, and Matplotlib, and provides a choice of Python and TensorFlow versions. From a hardware perspective, Google Colab is integrated with Google Drive and equipped with CPU, GPU, TPU, and RAM, enabling optimal processing performance as long as the internet connection is stable[11].

2.9. Naïve Bayes Classifier

The Naive Bayes Classifier was first proposed by British scientist Thomas Bayes using a probabilistic and statistical approach. This algorithm produces a simple probability-based prediction method by utilizing Bayes' Theorem. This means that in Naive Bayes, the model used is an "independent feature model." In Bayesian, particularly Naive Bayes, the strong independence of an element indicates that elements in the data are not related to the existence of other elements in the same data[12].

3. Research Method

3.1. CRISP-DM

In this study, the method used follows the CRISP-DM approach, which consists of several interrelated stages. The CRISP-DM stages are described as follows:

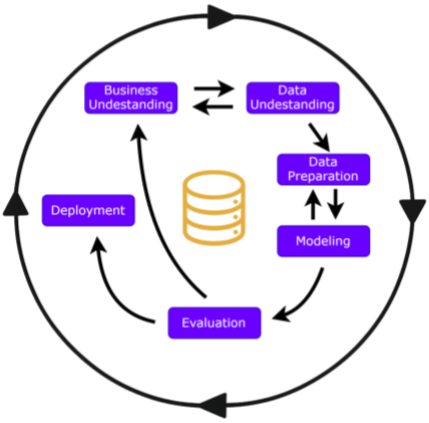

Fig. 1: Research Workflow

Based on the image above, it can be explained that there are several stages, as follows:

a) Business Understanding
At this stage, the research focuses on understanding the objective, which is to analyze student sentiment toward campus facilities and infrastructure. The analysis is conducted on comments collected from STMIK Kaputama students.

b) Data Understanding
This stage involves collecting student comments, followed by an initial exploration to identify sentiment distribution (positive, negative, and neutral) and assess data completeness.

c) Data Preparation
The collected comments are processed using text preprocessing, which includes cleaning, case folding, tokenizing, stopword removal, normalization, and stemming. The final result of this stage is clean text ready to be converted into numerical form using the TF-IDF method.

d) Modeling
The processed data is used to build a classification model with the Naïve Bayes Classifier algorithm. The dataset is divided into training data and test data, so that the model can learn from historical data and be tested with new data.

e) Evaluation
The model is evaluated using accuracy, precision, recall, and F1-score metrics. This evaluation aims to determine the model's performance in classifying student comments into three sentiment categories.

f) Deployment
The analysis results are presented in the form of reports and visualizations, such as confusion matrices, sentiment distributions, and word frequencies. This information is expected to serve as a basis for the university in improving its facilities and infrastructure.

3.2. Supporting Data

The data used in this study consists of 1266 comments from STMIK Kaputama students regarding campus facilities and infrastructure, such as classrooms, laboratories, libraries, toilets, parking areas, and internet access. These comments were collected in text form and became the basis for sentiment analysis. Some of the data is shown in the table below as examples of student comments.

Table1: Sample Comment Data

Comments
Parkiran sangat sempit, toilet bau
mungkin untuk kursi nya sih harus diganti lagi, karena yg lama sudah ada yg mejanya turun, injakan kakinya lepas dan lainnya.
untuk kebersihan kampus sudah jauh lebih baik dari tahun sebelumnya
Masih ada beberapa ruangan yang ac ya rusak dan juga wifi yg tidak lancar dan beberapa komputer yg tidak menyala
Toilet yang kurang bersih, parkir yang sempit dan tidak luas, proyek banyak yang rusak, kursi banyak yang rusak, akses internet yang susah diakses.

4. Results and Discussion

4.1. Text Preprocessing Results

Student comments were processed through several stages, namely cleaning, case folding, tokenizing, stopword removal, normalization, and stemming. This process produced cleaner, more uniform text that was ready to be converted into numerical form. The following are the results of text preprocessing.

Table 2: Text preprocessing results

Before Text Preprocessing	After Text Preprocessing
Parkiran sangat sempit, toilet bau	parkir sempit toilet bau
mungkin untuk kursi nya sih harus diganti lagi, karena yg lama sudah ada yg mejanya turun, injakan kakinya lepas dan lainnya.	kursi sih ganti meja turun injak kaki lepas
untuk kebersihan kampus sudah jauh lebih baik dari tahun sebelumnya	Bersih kampus
Masih ada beberapa ruangan yang ac ya rusak dan juga wifi yg tidak lancar dan beberapa komputer yg tidak menyala	ruang ac rusak wifi lancar komputer nyala
Toilet yang kurang bersih, parkir yang sempit dan tidak luas, proyek banyak yang rusak, kursi banyak yang rusak, akses internet yang susah diakses.	toilet bersih parkir sempit luas proyek rusak kursi rusak akses internet susah akses

4.2. Data Representation (Split Data and TF-IDF)

The processed comment data was divided into 976 training set comments and 244 testing set comments. Next, the data was converted into numerical representation using the TF-IDF method, which assigns a weight to each word based on its level of importance in the document.

Table 3: Example of TF-IDF results

Frekuensi					TF (t,d)					IDF	TF-IDF				
D1	D2	D3	D4	D5	D1	D2	D3	D4	D5		D1	D2	D3	D4	D5
1	0	0	0	1	0.250	0	0	0	0.077	0.398	0.099	0	0	0	0.031
1	0	0	0	1	0.250	0	0	0	0.077	0.398	0.099	0	0	0	0.031
1	0	0	0	1	0.250	0	0	0	0.077	0.398	0.099	0	0	0	0.031
1	0	0	0	0	0.250	0	0	0	0	0.699	0.175	0	0	0	0

4.3. Model Evaluation Results

The model was evaluated using 244 comments as test data. The test aimed to determine the extent to which the Naïve Bayes Classifier model was able to classify student comments into positive, negative, and neutral sentiment categories.

From the model testing results, the accuracy obtained was 0.73 or 73%, indicating that the model can predict correctly on most of the test data. The positive class had the best performance with precision of 0.72, recall of 0.97, and f1-score of 0.82, indicating a high ability to recognize positive comments. Conversely, the negative class still performed poorly with precision of 0.79, recall of 0.38, and F1-score of 0.51, while the neutral class was not detected at all (all metrics were 0.00). These results indicate that the model is more dominant in predicting positive sentiment but struggles to distinguish between negative and neutral comments.

Akurasi Model: 0.73

Laporan Klasifikasi:

	precision	recall	f1-score	support
negatif	0.79	0.38	0.51	69.00
netral	0.00	0.00	0.00	19.00
positif	0.72	0.97	0.82	156.00
accuracy	0.73	0.73	0.73	0.73
macro avg	0.50	0.45	0.44	244.00
weighted avg	0.68	0.73	0.67	244.00

Fig.2: Model Evaluation

In addition, the distribution of sentiment labels is also visualized to see the amount of data in each category. Positive sentiment dominates with around 815 data points, followed by negative sentiment with 315 data points, while neutral sentiment is the least with 95 data points.

This imbalance in distribution indicates that the data is more focused on positive comments, making the model more likely to recognize patterns in that class. This condition also implies poor model performance in detecting negative or neutral comments, as the limited amount of training data results in weaker representation of these two classes compared to the positive class.

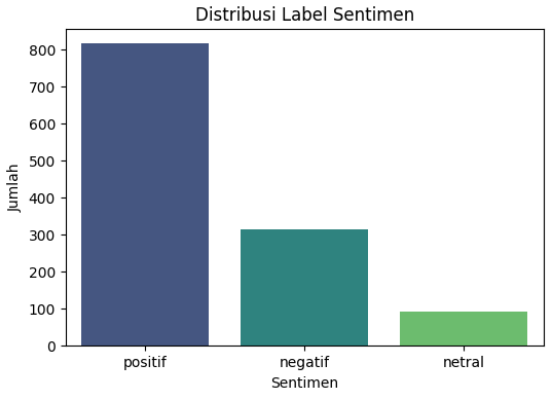

Fig.3: Sentiment Label Distribution

5. Conclusion and Recommendations

5.1. Conclusion

Based on the results of research conducted on the sentiment analysis of STMIK Kaputama student comments using the Naïve Bayes Classifier algorithm, the following conclusions were obtained:

1) The text preprocessing process, which includes case folding, tokenizing, stopword removal, and stemming, successfully converted the raw data into a more structured form for analysis.
2) Data representation using the TF-IDF method successfully converted the text comments into numerical vectors, enabling them to be processed by the classification algorithm.
3) The Naïve Bayes Classifier model achieved an accuracy of 73%, with the best performance in the positive class (precision 0.72, recall 0.97, F1-score 0.82). The negative class remains low (precision 0.79, recall 0.38, F1-score 0.51), while the neutral class is not detected at all (all metrics are 0.00).
4) The imbalanced data distribution, where positive comments are far more dominant than negative and neutral ones, affects the model's performance, which tends to more easily recognize the positive class.

5.2. Recommendations

Some suggestions for further research development are as follows:

1) Use a dataset with a more balanced distribution so that the model can learn more optimally across all sentiment classes.
2) Applying oversampling or undersampling techniques to address the imbalanced data problem.
3) Comparing the performance of the Naïve Bayes Classifier algorithm with other methods, such as Support Vector Machine (SVM) or Random Forest, to obtain more accurate classification results.
4) Integrating this sentiment analysis into the campus information system to support decision-making in improving the quality of facilities and services.

References

[1] R. Indonesia, “Undang-Undang Republik Indonesia Nomor 20 Tahun 2003 tentang Sistem Pendidikan Nasional,” Jakarta, 2003. [Online]. Available: https://peraturan.bpk.go.id/Details/43920/uu-no-20-tahun-2003
[2] S. Erisa, A. Sihotang, K. U. Almas, S. Mardiah, and D. A. Zahara, “Pengaruh Sarana Dan Prasarana Akademik Terhadap Minat Belajar Mahasiswa Pendidikan Ekonomi Stambuk 2023,” *J. EK&BI*, vol. 7, no. 1, pp. 48–56, 2024, doi: 10.37600/ekbi.v7i1.1338.
[3] T. N. Prakash and A. Aloysius, “Textual Sentiment Analysis using Lexicon Based Approaches,” *Ann. Rom. Soc. Cell Biol.*, vol. 25, no. 4, pp. 9878–9885, 2021, [Online]. Available: http://annalsofrscb.ro/index.php/journal/article/view/3734
[4] C. F. Hasri and D. Alita, “Penerapan Metode Naã□Ve Bayes Classifier Dan Support Vector Machine Pada Analisis Sentimen Terhadap Dampak Virus Corona Di Twitter,” *J. Inform. dan Rekayasa Perangkat Lunak*, vol. 3, no. 2, pp. 145–160, 2022, doi: 10.33365/jatika.v3i2.2026.
[5] N. Nurwanda, N. Suarna, and W. Prihartono, “Penerapan Nlp (Natural Language Processing) Dalam Analisis Sentimen Pengguna Telegram Di Playstore,” *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 2, pp. 1841–1846, 2024, doi: 10.36040/jati.v8i2.8469.
[6] A. Firdaus, W. I. Firdaus, P. Studi, T. Informatika, M. Digital, and P. N. Sriwijaya, “Text Mining,” vol. 13, no. 1, pp. 66–78, 2021.
[7] F. Lubis *et al.*, “Penggunaan Metode Text Mining Untuk Mengekstrak Informasi Penting Dari Teks Laporan Penelitian,” *J. Motiv. Pendidik. dan Bhs.*, vol. 1, no. 4, 2023, [Online]. Available: https://doi.org/10.59581/jmpb-widyakarya.v1i4.1961
[8] P. Yohana, S. Agustian, and S. K. Gusti, “Klasifikasi Sentimen Masyarakat terhadap Kebijakan Vaksin Covid-19 pada Twitter dengan Imbalance Classes Menggunakan Naive Bayes,” *Semin. Nas. Teknol. ...*, pp. 69–80, 2022, [Online]. Available: http://ejournal.uin-suska.ac.id/index.php/SNTIKI/article/view/19012%0Ahttp://ejournal.uin-suska.ac.id/index.php/SNTIKI/article/viewFile/19012/8336

[9] M. H. Mahendra, D. T. Murdiansyah, and K. M. Lhaksmana, "Analisis Sentimen Tweet COVID-19 menggunakan K-Nearest Neighbors dengan TF-IDF dan Ekstraksi Fitur CountVectorizer," *DIKE J. Ilmu Multidisiplin*, vol. 1, no. 2, pp. 37–43, 2023, doi: 10.69688/dike.v1i2.35.

[10] P. P. O. Mahawardana, I. A. P. F. Imawati, and I. W. Dika, "Analisis Sentimen Berdasarkan Opini dari Media Sosial Twitter terhadap 'Figure Pemimpin' Menggunakan Python," *J. Manaj. dan Teknol. Inf.*, vol. 12, no. 2, pp. 50–56, 2022, [Online]. Available: https://ojs.mahadewa.ac.id/index.php/jmti/article/view/2111

[11] R. T. Handayanto and H. Herlawati, "Prediksi Kelas Jamak dengan Deep Learning Berbasis Graphics Processing Units," *J. Kaji. Ilm.*, vol. 20, no. 1, pp. 67–76, 2020, doi: 10.31599/jki.v20i1.71.

[12] F. Harahap, N. E. Saragih, E. T. Siregar, and H. Sariangsah, "Penerapan Data Mining Dengan Algoritma Naive Bayes Classifier Dalam Memprediksi Pembelian Cat," *J. Ilm. Inform.*, vol. 9, no. 01, pp. 19–23, 2021, doi: 10.33884/jif.v9i01.3702.

[13] Muammar Khadapi, & Pakpahan, V. M. (2024). Analisis Sentimen Berbasis Jaringan LSTM dan BERT terhadap Diskusi Twitter tentang Pemilu 2024. *JUKI : Jurnal Komputer Dan Informatika*, *6*(2), 130–137. Retrieved from https://ioinformatic.org/index.php/JUKI/article/view/681