JoCoSiR
Journal of Computer Science Research
ISSN 2986-2337 (Online)

Sentiment Analysis of Indonesian TikTok Comments Using TF-IDF with Naive Bayes and SVM

Rezkinah Rambe [a], Muhammad Iqbal [b]

[a,b] Magister Teknologi Informasi, Universitas Pembangunan Panca Budi, JL. Gatot Subroto, Kec. Medan Sunggal, Kota Medan, Sumatera Utara, Indonesia

email: [a] renarrambe@gmail.com, [b] wakbalpb@yahoo.co.id

ARTICLE INFO

Keywords:
sentiment analysis,
TikTok comments,
Indonesian language,
TF-IDF,
Naive Bayes,
SVM.

IEEE style in citing this article:
R. Rambe and M. Iqbal, " Sentiment Analysis of Indonesian TikTok Comments Using TF-IDF with Naive Bayes and SVM," *JoCoSiR: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 3, no. 02, pp. 58-61, 2025.

ABSTRACT

This study aims to develop an automatic sentiment classification model for Indonesian TikTok comments using Term Frequency–Inverse Document Frequency (TF-IDF) with Naive Bayes and Support Vector Machine (SVM). Fifteen thousand comments were collected from public TikTok videos and manually labeled as positive, negative, and neutral. Data preprocessing included case folding, tokenization, stopword removal, and stemming (Nazief-Adriani algorithm). TF-IDF weighting transformed text into vectors, then used to train both classifiers. Performance was evaluated using accuracy, precision, recall, and F1-score trough 5-fold cross-validation. Result show SVM outperforms Naive Bayes with 92.8% accuracy versus 83%. Findings confirm that TF-IDF combined with SVM produces more relieble result for short Indonesian text classification, offering valuable insights for social media monitoring applications.

Copyright: Journal of Computer Science Research (JoCoSiR) with CC BY NC SA license.

1. Introduction

TikTok has become one of Indonesia's most dynamic and widely used social-media platforms, producing millions of user interactions and comments each day. These comments contain valuable insights into public perception, emotional reactions, and collective opinions toward various forms of content, including entertainment videos, promotional material, and public-policy discussions. However, the linguistic characteristics of TikTok comments pose challenges for analysis. Users frequently employ informal language, slang, sarcasm, emojis, abbreviations, and a mixture of Indonesian, English, and regional dialects. As a result, manually reviewing and interpreting these comments is time-consuming, inconsistent, and prone to subjective bias, which limits its effectiveness as a large-scale analytical approach.

Sentiment analysis provides a computational method for addressing this challenge by leveraging Natural Language Processing (NLP) and machine learning techniques to automatically classify the emotional tone expressed in text. By identifying whether a comment conveys positive, negative, or neutral sentiment, researchers and organizations can gain structured insight into audience reactions. Nevertheless, the effectiveness of sentiment analysis depends heavily on algorithm selection, preprocessing strategies, and feature-representation techniques. While previous research has explored sentiment classification on platforms such as Twitter, Instagram, and YouTube, studies focusing specifically on Indonesian TikTok comment patterns remain relatively scarce, leaving a methodological gap that needs to be addressed.

This study aims to bridge that gap by proposing a comparative framework to evaluate the performance of two widely used classification algorithms: Naive Bayes and Support Vector Machine (SVM). Both algorithms are implemented using TF-IDF (Term Frequency–Inverse Document Frequency) to convert text into weighted numerical features. Naive Bayes offers advantages in computational efficiency and simplicity, particularly for sparse text data, while SVM is known for robustness in handling high-dimensional data and producing more accurate decision boundaries. By applying these algorithms to real TikTok comment data, this study examines how well each model adapts to the informal and variable linguistic structures typical of Indonesian social-media communication.

The objective of this research is to determine the most effective sentiment-classification approach for Indonesian TikTok comments, particularly in multi-class classification scenarios. The results are expected to provide practical insights for researchers, marketing analysts, digital-platform developers, and policymakers

ISSN 2502-3357 (Online)

seeking to understand public opinion in real-time. Moreover, the findings may serve as a foundation for future development of more advanced models, such as deep learning architectures and transformer-based language models tailored to Indonesian social-media contexts. Ultimately, this study contributes to the broader field of sentiment analysis by demonstrating how algorithmic approaches can be optimized for short, informal, and highly dynamic user-generated text.

2. Literature Review

2.1 Sentiment Analysis

Sentiment analysis explores public emotions expressed through text by categorizing opinions as positive, n egative, or neutral. Researchers such as Liu (2012) define it as an NLP process to extract subjectivity and polarity f rom unstructured data. Applications include product-review analysis and social monitoring.

2.2 Term Frequency–Inverse Document Frequency (TF-IDF)

TF-IDF is a numerical statistic that reflects the importance of a term in a document relative to a corpus. A hi gh TF-IDF value indicates that a word is rare yet significant in distinguishing content. This technique is commonly combined with machine-learning classifiers for text tasks.

$$TF\ -\ IDF(t,d) = TF(t,d) \times \log\left(\frac{N}{DF(t)}\right)$$

2.3 Naive Bayes Classifier

Naive Bayes is a probabilistic model based on Bayes' theorem that assumes independence between features . Multinomial Naive Bayes is effective for word-count features and computationally efficient for large datasets:

$$P(C|X) \propto P(C)\prod_{i=1}^{n} P(x_i|C)$$

2.4 Support Vector Machine (SVM)

SVM is a supervised learning model that separates data points with an optimal hyperplane, maximizing the margin between classes. It works well with high-dimensional data like TF-IDF matrices.

$$\min_{w,b}\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i, y_i(w^T x_i + b) \geq 1 - \xi_i$$

3. Method

3.1 Research Design

The research flow includes six core stages: data collection, annotation, preprocessing, feature extraction, m odel training, and performance evaluation.

Crawling → Labeling → Preprocessing → TF-IDF Vectorization → Training (NB & SVM) → Evaluation

3.2 Data Collection

Data were scraped from TikTok comments using Python Selenium and BeautifulSoup between July and Se ptember 2025. Collected dataset: 15,000 comments in Indonesian. Each record contains a username, comment text , and timestamp. Example raw comment: *"Videonya keren banget!!! #viral 😂"*

3.3 Annotation

Three annotators classified comments as:

1. Positive: praise or approval (e.g., *"suka banget"*)
2. Negative: criticism or complaint (e.g., *"gak lucunya kelewatan"*)
3. Neutral: factual or question (e.g., *"kapan upload lagi?"*)

Majority-voting produced final labels. Inter-annotator agreement measured via Cohen's Kappa = 0.78.

3.4 Preprocessing

Steps applied to text data:

Step	Description	Example
Case Folding	lowercase conversion	"Bagus" → "bagus"
Tokenization	sentence splitting into tokens	"video bagus banget" → [video, bagus, b anget]
Cleaning	remove numbers, emoji, URL, punctu ation	"keren 😂 https://" → "keren"
Stopword Remo val	remove common function words	"yang", "dan"
Stemming	apply Nazief-Adriani algorithm	"menyukai" → "suka"

After preprocessing, vocabulary reduced ≈ 75 %.

3.5 Feature Extraction

DOI: https://doi.org/10.65126/jocosir.v3i2
Journal of Computer Science Research (JoCoSiR) with CC BY NC SA license

ISSN 2502-3357 (Online)

TF-IDF was implemented in *Scikit-learn* with:

max_features = 5000
ngram_range = (1, 2)
min_df = 2 , max_df = 0.8
sublinear_tf = True

Each comment was represented as a 5,000-dimensional vector.

3.6 Dataset Split

Data were split into:

1. 80 % training (12,000 comments)
2. 20 % testing (3,000 comments)

Stratified split was used to maintain class balance.

3.7 Model Implementation

1. Naive Bayes
   from sklearn.naive_bayes import MultinomialNB
   NB = MultinomialNB(alpha=1.0)
   NB.fit(X_train, y_train)
2. Support Vector Machine
   from sklearn.svm import LinearSVC
   SVM = LinearSVC(C=1.0, max_iter=1000)
   SVM.fit(X_train, y_train)

Evaluation Metrics

Accuracy, Precision, Recall, and F1-Score calculated using Scikit-learn's classification_report.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}, \quad F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

3.8 Validation

5-Fold Cross-Validation ensured robustness of each model. Confusion matrices provided error distributions for positive, negative, neutral classes.

4. Results and Discussion

Model	Accuracy	Precision	Recall	F1-Score
Naive Bayes	83 %	0.83	0.82	0.82
SVM	92.8 %	0.92	0.93	0.92

Analysis

1. SVM performed better due to its capacity in handling high-dimensional sparse TF-IDF vectors.
2. Naive Bayes is faster but assumes feature independence, limiting accuracy.
3. Frequent errors involve sarcastic tones (e.g., *"keren banget bikin ngantuk"*) and code-mixed phrases.
4. Preprocessing and TF-IDF significantly improved accuracy (from 78 % to 93 %).

Computational Performance:

SVM training ≈ 18 seconds; NB ≈ 3 seconds. Memory usage increased for SVM but within tolerable limits.

The evaluation results show that the SVM model outperformed Naive Bayes in classifying sentiment in Indonesian TikTok comments. As presented in the performance metrics, SVM achieved an accuracy of 92.8%, while Naive Bayes reached 83%. This indicates that SVM is more effective in distinguishing sentiment patterns in short and informal text. The higher precision, recall, and F1-score of SVM further signify its ability to classify positive, negative, and neutral sentiments more consistently across different comment variations.

The superior performance of SVM in this study can be attributed to its ability to handle high-dimensional and sparse TF-IDF feature vectors, which are common in social-media text. TikTok comments often contain slang, abbreviations, emojis, and mixed language usage, resulting in a wide vocabulary with many rarely used terms. SVM is capable of constructing an optimal hyperplane that separates sentiment classes in such conditions. In contrast, Naive Bayes assumes independence between features, which limits its ability to understand contextual meaning and subtle emotional cues, leading to lower accuracy in complex linguistic patterns.

Both models showed misclassification tendencies in comments containing sarcasm or code-mixing. Phrases such as "keren banget bikin ngantuk" or "mantap lah bos, kecewa kali pun" may appear positive in literal terms but actually convey negative sentiment. Such expressions require models capable of understanding pragmatic and contextual interpretation beyond lexical features alone. This suggests that future research could explore transformer-based language models (e.g., IndoBERT), which are better suited for capturing contextual semantics in Indonesian social-media discourse.

DOI: https://doi.org/10.65126/jocosir.v3i2
Journal of Computer Science Research (JoCoSiR) with CC BY NC SA license

ISSN 2502-3357 (Online)

Preprocessing and the use of TF-IDF significantly improved model performance, raising accuracy from an initial 78% to 93%. Cleaning the text by removing noise, standardizing word forms, and converting text to weighted numeric features helped reduce ambiguity and highlight meaningful patterns. From a computational perspective, Naive Bayes trained faster (≈ 3 seconds) and used less memory compared to SVM (≈ 18 seconds). However, the additional computational cost of SVM remains within acceptable limits and is justified by the substantial gain in classification accuracy. Therefore, SVM is more suitable for sentiment analysis applications where precision and reliability are prioritized, while Naive Bayes is appropriate when speed and low resource usage are the primary considerations.

5. Conclusions

This research successfully developed a TikTok comment sentiment-analysis pipeline using TF-IDF vectori zation and two machine-learning algorithms. SVM achieved the highest accuracy (92.8 %), while Naive Bayes ach ieved 83 %. Text preprocessing was found to be critical for accuracy improvement.
Future research should explore contextual embeddings (IndoBERT) and multilingual handling for slang and sarcas m detection.

6. References

[1] Arsyah, U. I., et al. (2024). *Twitter Sentiment Analysis of Public Opinions Using SVM and TF-IDF*. IJCS.

[2] Aispurs, V. (2024). *Exploring Sentiment Analysis on TikTok Social Media Platform.* Univ. Stockholm.

[3] Prihatini, P. M. (2023). *Feature Extraction for Indonesian Text Classification*. Proc. ICAS.

[4] Ariyus, D., et al. (2024). *Enhancing Sentiment Analysis of TikTok Comments Using FastText and Bi-LSTM.* ETASR.

[5] Pebriana, S., & Sugianto, C. A. (2025). *Sentiment Analysis of Netizen Opinions on TikTok Toward iPhone.* J -AI SE. https://doi.org/10.30811/jaise.v5i3.7011

[6] Rahmadani, P. S. (2022). *Tiktok Social Media Sentiment Analysis Using Naïve Bayes Classifier.* Sinkron Jou rnal, 7 (3), 995-999. https://doi.org/10.33395/sinkron.v7i3.11579

[7] Faddilla A. D. & Pratama I. (2025). *Sentiment Analysis of TikTok Tokopedia Seller Center Application Usin g SVM and Naive Bayes.* IJSECS. https://doi.org/10.35870/ijsecs.v5i1.3463

[8] Hidayah, I. A., Kusumawati, R., Abidin, Z., & Imamudin, M. (2024). Analysis of public sentiment towards the Tiktok application using the Naive Bayes Algorithm and support Vector machine. *Journal of Computer Networks, Architecture and High Performance Computing*, *6*(2), 881-891. https://doi.org/10.47709/cnahpc.v6i2.3990

[9] Silitonga, P. D., Hasibuan, M., Situmorang, Z., & Purba, D. (2023). Comparison of TikTok user sentiment analysis accuracy with Naïve Bayes and support vector machine. *International Journal*, *12*(1).

[10] Aji, S., Sundari, J., Yunita, Imron, & Pratama, O. (2023, May). The algorithm comparison of support vector machine and Naive Bayes in sentiment analyzing the Tiktok application. In *2ND INTERNATIONAL CONFERENCE ON ADVANCED INFORMATION SCIENTIFIC DEVELOPMENT (ICAISD) 2021: Innovating Scientific Learning for Deep Communication* (Vol. 2714, No. 1, p. 020015). AIP Publishing LLC. https://doi.org/10.1063/5.0129009

[11] Akbar, M. R., & Defit, S. (2024). Metode Support Vector Machine dan Naïve Bayes Untuk Analisis Santimen Ibu Kota Nusantara. *Jurnal KomtekInfo*, 323-331. https://doi.org/10.35134/komtekinfo.v11i4.579

[12] Ariyus, D., Manongga, D., & Sembiring, I. (2024). Enhancing Sentiment Analysis of Indonesian Tourism Video Content Commentary on TikTok: A FastText and Bi-LSTM Approach. *Engineering, Technology & Applied Science Research*, *14*(6), 18020-18028. https://doi.org/10.48084/etasr.8859

[13] Rahmadani, P. S., Tampubolon, F. C., Jannah, A. N., Hutabarat, N. L. H., & Simarmata, A. M. (2022). Tiktok Social Media Sentiment Analysis Using the Nave Bayes Classifier Algorithm. *Sinkron: jurnal dan penelitian teknik informatika*, *6*(3), 995-999. 10.33395/sinkron.v7i3.11579

[14] Akbar, M. R., & Defit, S. (2024). Metode Support Vector Machine dan Naïve Bayes Untuk Analisis Santimen Ibu Kota Nusantara. *Jurnal KomtekInfo*, 323-331.https://doi.org/10.35134/komtekinfo.v11i4.579

[15] Aji, S., Sundari, J., Yunita, Imron, & Pratama, O. (2023, May). The algorithm comparison of support vector machine and Naive Bayes in sentiment analyzing the Tiktok application. In *2ND INTERNATIONAL CONFERENCE ON ADVANCED INFORMATION SCIENTIFIC DEVELOPMENT (ICAISD) 2021: Innovating Scientific Learning for Deep Communication* (Vol. 2714, No. 1, p. 020015). AIP Publishing LLC. https://doi.org/10.1063/5.0129009

DOI: https://doi.org/10.65126/jocosir.v3i2
Journal of Computer Science Research (JoCoSiR) with CC BY NC SA license