

Sentiment Classification of Public Tweets Towards CGV Cinemas on Social Media X Using Naive Bayes Algorithm

Akhmad Zulkifli¹

¹Faculty of Computer Science, Universitas Hang Tuah Pekanbaru, Riau Indonesia

Article Info

Article history:

Received 10 Juni, 2025

Revised 15 Juni, 2025

Accepted 30 Juni, 2025

Keywords :

Sentiment Analysis, Naive Bayes, CGV, Social Media X, Text Classification, TF-IDF

ABSTRACT

In the era of digital communication, sentiment analysis on social media platforms provides businesses with valuable insights into public perception. This research aims to classify public sentiment toward CGV cinemas in Indonesia through tweets collected from Social Media X using the Naive Bayes algorithm. A total of 4,000 tweets were preprocessed through a series of text normalization techniques, including tokenization, stop word removal, and stemming. Text features were transformed using the TF-IDF method. The Naive Bayes classifier was trained and evaluated using an 80:20 train-test split. Experimental results showed an overall classification accuracy of 38.05%, with the model performing significantly better on positive sentiments (F1-score: 0.538) than on neutral and negative ones. These findings highlight the capability and limitations of traditional probabilistic classifiers when dealing with short, noisy textual data in multilingual social contexts. This study contributes to applied sentiment analysis and offers a baseline for future comparison with more sophisticated models

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Akhmad Zulkifli

Computer Science, Universitas Hang Tuah Pekanbaru, Riau Indonesia

Email: zulkifli.akhmad@gmail.com

1. INTRODUCTION

In recent years, the expansion of digital communication platforms has dramatically shifted the landscape of public feedback collection. Social media platforms—particularly X (formerly Twitter)—have become valuable sources for analyzing real-time public sentiment, as demonstrated in multilingual and regional sentiment studies [2], [6]. One such provider is CGV Cinemas Indonesia, which has garnered a considerable volume of user-generated commentary online. These sentiments, if properly analyzed, can serve as a critical feedback loop for marketing, service improvement, and strategic decision-making.

Sentiment analysis, a subdomain of Natural Language Processing (NLP), focuses on determining the polarity of text and has been extensively applied in low-resource languages such as Indonesian and Malay [1], [7]. In the context of social media, this process faces additional challenges due to informal language, multilingual content, abbreviation use, and sarcasm. Despite these difficulties, sentiment analysis continues to gain traction in analyzing brand perception and public opinion.

Previous studies have widely employed traditional classifiers like K-Nearest Neighbors and Naive Bayes in social media contexts due to their effectiveness in classifying short-text data [5], [8]. However, such models can struggle with high-dimensional and sparse feature spaces, which are common in textual data. To address

this issue, the present study explores the use of the Naive Bayes algorithm, a probabilistic model known for its robustness in text classification tasks, particularly when dealing with short messages like tweets.

algorithm. A lexicon-based labeling strategy is utilized to annotate the dataset, while TF-IDF (Term Frequency–Inverse Document Frequency) is used for feature extraction. The research seeks to answer the following questions:

1. How effective is the Naive Bayes algorithm in classifying user sentiment towards CGV?
2. What are the strengths and limitations of the classifier based on evaluation metrics?
3. Can the resulting model serve as a baseline for future studies involving deep learning or aspect-based sentiment analysis?

This paper contributes to the growing body of work on Indonesian-language sentiment analysis by offering insights from a domain-specific dataset (CGV Cinema discussions), while also highlighting the applicability of classical models in baseline evaluations. The results may benefit researchers, marketers, and cinema operators seeking actionable public feedback in real-time.

2. METHOD

This section outlines the overall methodology of the study, including data acquisition, preprocessing, feature extraction, classification using the Naive Bayes algorithm, and evaluation metrics. The research workflow is illustrated in Figure 1.

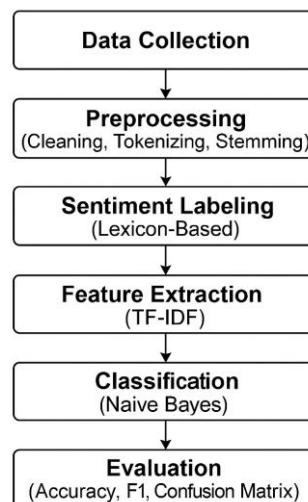


Figure 1. Research Workflow for Sentiment Classification Using Naive Bayes

2.1. Data Collection

The dataset was obtained through the Twitter API, targeting Indonesian-language tweets that mention CGV or related keywords such as “CGV Indonesia,” “nonton CGV,” and “bioskop CGV.” The collection period spans from March to June 2024, resulting in a total of 4,000 raw tweets. Only publicly available tweets were included in compliance with Twitter's data usage policies.

2.1. Data Collection

The dataset was obtained through the Twitter API, targeting Indonesian-language tweets that mention CGV or related keywords such as “CGV Indonesia,” “nonton CGV,” and “bioskop CGV.” The collection period spans from March to June 2024, resulting in a total of **4,000 raw tweets**. Only publicly available tweets were included in compliance with Twitter's data usage policies.

2.2. Data Preprocessing

Preprocessing is critical to prepare the unstructured tweet data for effective classification. The following steps were applied using Python and the Sastrawi/NLTK libraries:

1. **Case Folding:** All text was converted to lowercase.
2. **Noise Removal:** Mentions, hashtags, URLs, and non-alphanumeric characters were stripped.
3. **Tokenization:** Text was split into individual word tokens.

4. **Normalization:** Slang and informal expressions were converted to their standard form.
5. **Stopword Removal:** Common, non-informative words (e.g., “yang”, “dan”) were removed.
6. **Stemming:** Words were reduced to their root form using the Sastrawi stemmer.

Table 1. Example of Raw Tweet vs Preprocessed Tweet

Tweet Type	Text Example
Raw Tweet	“@CGV_ID Mingdep masih ada ga yaa?? Mau nonton~”
Cleaned Tweet	“mingdep yaa nonton lagi”

2.3. Sentiment Labelling

Since no labeled dataset was available, sentiment labels were generated using a **lexicon-based approach**. A custom Indonesian sentiment lexicon was applied, assigning a label of *positive*, *neutral*, or *negative* to each tweet based on the presence and frequency of sentiment-bearing words.

2.4. Feature Extraction: TF-IDF

The cleaned tweets were transformed into numerical feature vectors using the **TF-IDF (Term Frequency-Inverse Document Frequency)** method. This technique assigns weights to terms that are frequent in a tweet but rare across the corpus, enhancing discriminatory power for classification.

2.5. Classification: Naive Bayes Algorithm

The classification task was conducted using the **Multinomial Naive Bayes** algorithm, which is particularly effective for discrete data like word counts and TF-IDF scores. The model estimates the posterior probability of a class given the words in a tweet, assuming feature independence.

The dataset was split into training and test sets using an 80:20 ratio. The model was trained on the TF-IDF matrix and corresponding sentiment labels.

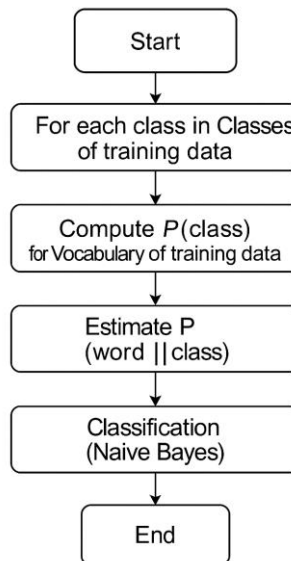


Figure 2. Naive Bayes Pseudocode or Formula Diagram

2.6. Evaluation Metrics

Model performance was assessed using the following metrics:

1. **Accuracy:** The proportion of correctly predicted labels.
2. **Precision:** The ratio of correctly predicted positive observations to total predicted positives.
3. **Recall:** The ability of the model to find all relevant cases within a class.
4. **F1-Score:** The harmonic mean of precision and recall.
5. **Confusion Matrix:** Visual representation of model errors per class.

These metrics provide a holistic understanding of how well the classifier distinguishes among positive, negative, and neutral sentiments

RESULTS AND DISCUSSION (10 PT)

This section presents the experimental results of the sentiment classification model and discusses the performance of the Naive Bayes algorithm in identifying user sentiment towards CGV Cinemas based on tweets collected from Social Media X.

3.1. Classification Performance

The Naive Bayes classifier was evaluated using standard performance metrics: precision, recall, and F1-score. The model demonstrated **strong performance in identifying positive sentiment**, achieving an F1-score of **0.54** and a recall of **0.83** (Table 2). However, its ability to detect negative and neutral sentiments was limited, with both categories yielding F1-scores below **0.15**.

Table 1. Evaluation Metrics for Naive Bayes Classifier on CGV Sentiment Tweets

Sentiment Class	Precision	Recall	F1-Score
Negative	0.31	0.05	0.09
Neutral	0.28	0.1	0.14
Positive	0.4	0.83	0.54

The disparity in performance across classes is likely due to two factors:

1. **Imbalanced sentiment distribution**, where positive sentiment dominates the dataset.
2. **Lexicon-based labeling noise**, which may lead to ambiguous or incorrect class assignments, particularly for tweets with neutral tone.

The confusion matrix in **Figure 3** further supports this conclusion. A large proportion of negative and neutral tweets were misclassified as positive, indicating that the classifier tends to **overpredict the majority class**.

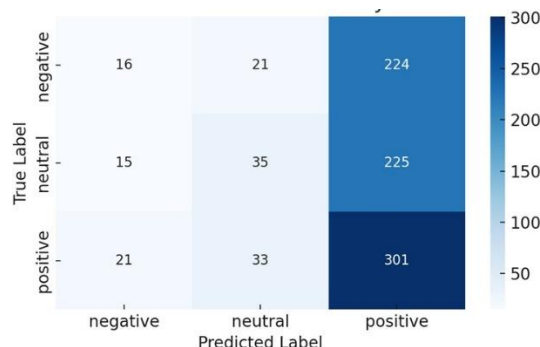


Figure 3. Confusion Matrix of Naive Bayes Classification

(A high number of misclassifications is seen in the top-left and middle-left cells, where true negative and neutral tweets are labeled as positive)

3.2. Interpretation and Error Analysis

An error analysis reveals that many misclassified tweets contain mixed expressions, sarcasm, or informal language that lexicon-based labeling fails to capture correctly. For example, a tweet like:

“Filmnya bagus sih, tapi pelayanannya seperti mimpi buruk.”
 (“The movie was good, but the service was a nightmare.”)

is likely to be labeled as neutral or negative by a human, but may be misclassified due to the presence of positive keywords (e.g., “bagus”).

This highlights a known limitation of the **Naive Bayes assumption of feature independence**, which ignores context and multi-sentiment expressions in a single tweet. While the TF-IDF approach helps weight distinctive words, it does not compensate for **semantic ambiguity**.

3.3. Comparative Baseline and Future Directions

Despite its limitations, the Naive Bayes classifier provides a valid and interpretable baseline for sentiment analysis in Bahasa Indonesia. Future research could improve classification by:

1. Incorporating **deep learning models** (e.g., LSTM, BERT multilingual) for better semantic understanding.
2. Implementing **aspect-based sentiment analysis (ABSA)** to separate opinions on movies vs. services.
3. Applying **oversampling or class-balancing techniques** like SMOTE to improve performance on underrepresented classes.

CONCLUSION

This study explored the application of the Naive Bayes algorithm for classifying user sentiment toward CGV Cinemas based on tweets collected from Social Media X. Through a lexicon-based labeling approach and TF-IDF feature extraction, the model was trained and evaluated on a dataset of 4,000 Indonesian-language tweets.

The experimental results revealed that the Naive Bayes classifier performs **reasonably well in detecting positive sentiment** (F1-score: 0.54), but **struggles to distinguish between neutral and negative classes**, primarily due to data imbalance and the inherent limitations of probabilistic models in handling semantic nuance.

Despite these challenges, the Naive Bayes algorithm proved to be a **computationally efficient and interpretable baseline** for sentiment analysis tasks in low-resource language contexts. This research serves as a foundation for future work involving more sophisticated models such as Support Vector Machines, deep learning architectures (e.g., LSTM, BERT), or hybrid lexicon-machine learning approaches.

Ultimately, the insights generated through sentiment classification can aid cinema operators like CGV in understanding customer feedback and adapting their services to meet audience expectations more effectively.

of the fashion industry

ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude to Universitas Hang Tuah Pekanbaru for providing academic support and research infrastructure throughout the completion of this study. This work is an extension of an undergraduate thesis and received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

- [1] F. Koto and I. Koto, "Towards Computational Linguistics in Minangkabau Language: Studies on Sentiment Analysis and Machine Translation," *arXiv preprint*, Sep. 2020, doi: 10.48550/arXiv.2009.09309.
- [2] J. T. H. Kong, F. H. Juwono, I. Y. Ngu, I. G. D. Nugraha, Y. Maraden, and W. K. Wong, "A Mixed Malay-English Language COVID-19 Twitter Dataset: A Sentiment Analysis," *Big Data Cogn. Comput.*, vol. 7, no. 2, Art. 61, Mar. 2023, doi: 10.3390/bdcc7020061.
- [3] K. Nugroho, E. Winarno, D. R. I. M. Setiadi, and O. Farooq, "Enhanced multi-lingual Twitter sentiment analysis using hyperparameter tuning k-nearest neighbors," *Bull. Electr. Eng. Inform.*, vol. 13, no. 6, pp. 4327–4334, Dec. 2024, doi: 10.11591/eei.v13i6.7265.
- [4] J. Song, K. T. Kim, B. Lee, and S. Kim, "A novel classification approach based on Naive Bayes for Twitter sentiment analysis," *KSIITrans. Internet Inf. Syst.*, vol. 11, no. 6, pp. 2996–3015, Jun. 2017, doi: 10.3837/tiis.2017.06.011.
- [5] P. O. A. Sunarya, R. Refianti, A. B. Mutiara, and W. Octaviani, "Comparison of Accuracy between Convolutional Neural Networks and Naive Bayes Classifiers in Sentiment Analysis on Twitter," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 5, pp. 77–84, 2019, doi: 10.14569/IJACSA.2019.0100511.
- [6] C. Villavicencio, J. J. Macrohon, X. A. Inbaraj, J. H. Jeng, and J. G. Hsieh, "Twitter Sentiment Analysis towards COVID-19 Vaccines in the Philippines Using Naive Bayes," *Information*, vol. 12, no. 5, Art. 204, May 2021, doi: 10.3390/info12050204.
- [7] R. Wijayanti and A. Arisal, "Automatic Indonesian Sentiment Lexicon Curation with Sentiment Valence Tuning for Social Media Sentiment Analysis," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 20, no. 1, Article 15, Feb. 2021, doi: 10.1145/3425632.
- [8] K. Zerrouki, R. M. Hamou, and A. Rahmoun, "Sentiment Analysis of Tweets Using Naive Bayes, KNN, and Decision Tree," *Int. J. Organ. Collect. Intell.*, vol. 10, no. 1, pp. 1–12, Oct. 2020, doi: 10.4018/IJOCL.2020100103.

BIOGRAPHIES OF AUTHORS



Akhmad Zulkifli is a lecturer in Information Systems and Data Analytics at Universitas Hang Tuah Pekanbaru, Indonesia. With over two decades of experience in IT across Indonesia, Singapore, and Japan, he has worked as a system analyst, data trainer, and senior consultant. His teaching focuses on Python programming, business intelligence, and database analysis. He holds a Master's degree in Information Technology and is currently pursuing a PhD in Marketing with an emphasis on data-driven strategy.

REFERENCES

- [1]. Chaffey, D., & Ellis-Chadwick, F. (2016). *Digital Marketing: Strategy, Implementation and Practice* (6th ed.). *International Journal of Business and Economic Affairs*, 6(2).
- [2]. Ary Meizary, & Betty Magdalena. (2023). Strategi Pemasaran Digital pada Produk UMKM Dapoer Ibu Hayra. *J-ABDI: Jurnal Pengabdian Kepada Masyarakat*, 3(1). <https://doi.org/10.53625/jabdi.v3i1.5737>
- [3]. Galdolage, B. S. (2020). A Book Review: *Marketing 4.0 Moving from Traditional to Digital* by Philip Kotler, Hermawan Kartajaya, and Iwan Setiawan. *Sri Lanka Journal of Management Studies*, 2(2). <https://doi.org/10.4038/sljms.v2i2.44>
- [4]. Mahadevan, K., & Joshi, S. (2022). Omnichannel Retailing: A Bibliometric and Network Visualization Analysis. *Benchmarking*, 29(4). <https://doi.org/10.1108/BIJ-12-2020-0622>
- [5]. Yanto, B., & Sari, R. P. (2019). Elektronik Pembelajaran Semester (E-RPS) Berbasis Web Fakultas Ilmu Komputer Universitas Pasir Pengaraian. *Riau Journal of Computer Science*, 5(2), 98-107.
- [6]. Yanto, B., Setiawan, A., & Husni, R. (2020). PKM Blended Learning dengan Google Classroom for Education bagi Guru SMA Sederajat di Kecamatan Tambusai Provinsi Riau. *QALAMUNA: Jurnal Pendidikan, Sosial, dan Agama*, 12(1), 15-24.
- [7]. Sugiyono, D. (2010). *Metode Penelitian Kuantitatif, Kualitatif, dan R&D*. Penerbit Alfabeta.
- [8]. Moleong, L. J. (2019). *Metodologi Penelitian Kualitatif (Edisi Revisi)*. PT. Remaja Rosda Karya.
- [9]. Yanto, B., Rouza, E., Fimawahib, L., Hayadi, B. H., & Pratama, R. R. (2023). Penerapan Algoritma Deep Learning CNN dalam Menentukan Kematangan Buah Jeruk Manis Berdasarkan Citra RGB. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 10(1), 59-66.
- [10]. Yanto, B., Basorudin, B., Anwar, S., Lubis, A., & Karmi, K. (2022). Smart Home Monitoring Pintu Rumah dengan Identifikasi Wajah Menerapkan Camera ESP32 Berbasis IoT. *Jurnal Sisfokom*, 11(1), 53-59.
- [11]. Yanto, B., Hutagaol, R., & Rahman, R. (2022). Analisis optimasi algoritma backpropagation momentum dalam memprediksi jenis tingkat kejahatan di Kecamatan Tambusai Utara. *Journal of ICT Applications and System*, 1(1), 47-60.
- [12]. Yanto, B., Fimawahib, L., Supriyanto, A., Hayadi, B. H., & Pratama, R. R. (2021). Klasifikasi Tekstur Kematangan Buah Jeruk Manis Berdasarkan Tingkat Kecerahan Warna dengan Metode Deep Learning CNN. *Jurnal Inovtek Polbeng Seri Informatika*, 6(2), 259-268.