Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2025. Vol. 4. No. 2; e-ISSN: 2808-4519

Implementation of Naive Bayes in Sentiment Analysis of CapCut App Reviews on the Play Store

Oka Alvianto[1]*, Willy Prihartono[2], Fathurrohman[3]

[1,2,3] *STMIK IKMI Cirebon*
okaalvianto@gmail.com [1*]

Abstract

The CapCut video editing application has gained significant popularity among mobile users. This study aims to analyze user sentiment towards CapCut reviews on the Play Store using the Naive Bayes algorithm. User reviews were collected and preprocessed to clean and prepare the text for analysis. The Naive Bayes algorithm was employed to classify the reviews into positive and negative sentiment categories. Findings indicate that the majority of user reviews are positive, highlighting features such as ease of use, attractive visual effects, and the ability to share videos on social media. However, negative reviews were also identified, primarily criticizing issues like bugs, intrusive advertisements, and limitations in specific features. This research provides valuable insights into user sentiment towards CapCut, which can be utilized by developers to enhance application quality and user experience.

Keywords: sentiment analysis, Naive Bayes, CapCut, user reviews, Play Store

1. Introduction

In recent decades, rapid advancements in the field of informatics have significantly influenced various aspects of daily life. Information and communication technologies have transformed how we interact, work, and learn. Mobile applications have become an integral part of modern life, offering practical solutions for communication, entertainment, and productivity [1],[2]. Among these, video editing applications have gained popularity due to the increasing demand for content creation and sharing on social media platforms [3].

CapCut, a mobile-based video editing application, has emerged as a widely used tool because of its user-friendly interface and extensive features [4]. This application allows users to edit videos efficiently and share them directly on social media platforms. Despite its popularity, CapCut's growing user base has also resulted in diverse reviews on the Play Store, reflecting both positive feedback and criticisms regarding its functionality, advertisements, and occasional bugs [5]. These user reviews provide a valuable source of data for sentiment analysis to understand user preferences and concerns.

Sentiment analysis, a text mining technique, leverages machine learning to classify textual data into sentiment categories such as positive, negative, or neutral. Among the many algorithms available, Naive Bayes is favored for its simplicity and efficiency in text classification tasks[6]., [7]. Despite challenges in handling informal language and ambiguous sentiments, Naive Bayes has been widely used in analyzing user reviews on platforms such as the Play Store [8], [9], [10].

This research aims to implement the Naive Bayes algorithm to analyze sentiment in user reviews of the CapCut application. By identifying sentiment patterns, the study provides actionable insights for developers to improve the application's quality, address user concerns, and enhance the overall user experience.

2. Research method

This research implements the Naive Bayes algorithm to analyze user sentiment in reviews of the CapCut application available on the Google Play Store. The methodology consists of several stages, including data collection, preprocessing, sentiment classification, and evaluation.

2.1. Data Collection

Data was obtained through web scraping techniques using the Google Play Scraper library. A total of 1,000 user reviews were collected, representing a variety of sentiments, ranging from positive to negative. The collected data includes text reviews, star ratings, and metadata such as review dates. Reviews were filtered based on their length and relevance to ensure quality. Only reviews in Indonesian and English were included in the analysis, while duplicates and irrelevant entries were removed

2.2. Data Preprocessing

The preprocessing stage included several steps to prepare the text for analysis:

- **Case Folding**: Converting all text to lowercase to maintain consistency.
- **Tokenization**: Breaking down text into individual tokens (words).
- **Stopword Removal**: Eliminating common but uninformative words such as "and" or "the."
- **Stemming**: Reducing words to their root forms using the Sastrawi library for the Indonesian language.

2.3. Sentiment Classification

The Naive Bayes algorithm was applied to classify the reviews into two categories: positive and negative. Reviews with a star rating of 4–5 were labeled as positive, while those with ratings of 1–2 were categorized as negative. The Term Frequency-Inverse Document Frequency (TF-IDF) method was used to assign weights to tokens, emphasizing words that are significant within individual reviews but less common across the dataset.

2.4. Model Evaluation

The model's performance was evaluated using several metrics:

- **Accuracy**: The percentage of correctly classified reviews.
- **Precision**: The accuracy of the positive or negative predictions.
- **Recall**: The proportion of actual positive or negative reviews correctly identified.
- **F1-Score**: A harmonic mean of precision and recall to evaluate overall model effectiveness**.**

2.5. Tools and Frameworks

The Google Play Scraper library was used for data collection, while data preprocessing and analysis were conducted using Python. Libraries such as Pandas, NumPy, and NLTK were utilized for text processing, and the Naive Bayes model was implemented using Scikit-learn.By following this methodology, the research aims to provide a comprehensive sentiment analysis of user reviews, offering actionable insights for application improvement.

3. Result and Discusion

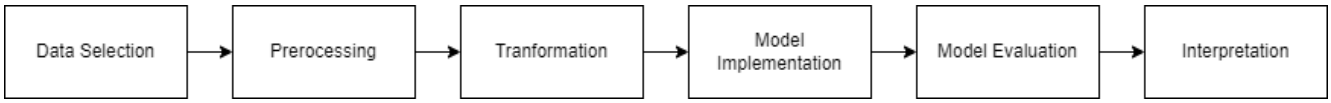

Fig. 1: KDD Diagram

This study aims to analyze the sentiment of user reviews for the CapCut application on the Google Play Store. The analysis involves several stages, including data collection, preprocessing, implementation of the Naive Bayes model, evaluation of results, and interpretation of findings. The steps taken and the outcomes achieved are detailed in this research.

3.1. Result

This research aims to analyze the sentiment of user reviews for the CapCut application available on the Google Play Store. The analysis was conducted through a series of stages, including data collection, preprocessing, implementation of the Naive Bayes model, evaluation of results, and interpretation of findings. The steps taken and the results obtained are described in this study.

3.1.1. Selection Data

The initial step in this research was data selection. User reviews of the CapCut application were obtained from the Google Play Store using web scraping techniques. The collected data included review text, star ratings, and other metadata information. To ensure the quality of the analysis, reviews that were too short, irrelevant, or contained only symbols were excluded. The data was then filtered to include various sentiment categories, namely positive and negative.

Table 1: Tabel Data Collection

Ulasan	Score	Label
dulu bagus banget,sekarang sering nge lag,terus kadang foto udah di edit tiba" hilang,kalo buat jj sering lambat gak mengikuti irama lagu nya	1	Negatif
agan di download!!apa apa harus pro dulu,buat video tugas malah jadi gagal,harga nya juga mahal buat di kantong pelajar.trus banyak iklan padahal dulu gak.sekarang juga ngelag banget gak jelas gak kek kayak dulu,mending gitu uninstall aja	1	Negatif
malah keluar! Itu membuat saya emosi!!jadi nya g nyaman! Tolong update capcut yang kayak dulu! Lebih enak yang dulu! Apalagi sekarang malah ada fitur pro nya segala! Kalo ada orang yang ga mampu gimana?? Yang bagus bagus di pro in teruss, capee tau, sekali nya ada pro tapi yang free Tolong perbaiki lagi bug nya dan iklan nyaa! Apalagi saya sudah ekspor malah keluar! Itu membuat saya emosi!!jadi nya g nyaman! Tolong update capcut yang kayak dulu! Lebih enak yang dulu! Apalagi sekarang malah ada fitur pro nya segala! Kalo ada orang yang ga mampu gimana?? Yang bagus bagus di pro in teruss, capee tau, sekali nya ada pro tapi yang free iklan dulu mana sampe 5/5 kali nontonin iklan lagi! Yaa walau bagus bagus sihh, tapi ga gini juga!	5	Positif

dulu bagus, tapi sekarang makin banyak iklan durasinya 30detik bahkan muncul berkali kali. tiba² suka masuk ke website. lemot banget, mau pake efek susah ga muncul². harus relog dulu muncul. padahal jaringan bagus.!!!! saat mau ngedit mau nambahkan lagu ngaturnya agak susah gak jelas!!, pokonya sekarang jelek!!!	1	Negatif

3.1.2. Preprocessing Data

The preprocessing process was carried out to clean and prepare the data for sentiment analysis. The steps involved include:

- **Case Folding**

  Case folding is the initial step where all text is converted to lowercase. The purpose is to standardize the format so that there is no distinction between uppercase and lowercase letters that could affect the analysis results. For example, the words "CapCut" and "capcut" will be considered identical after the case folding process.

Table 2: Case Folding

Ulasan	Case Folding
dulu bagus banget,sekarang sering nge lag,terus kadang foto udah di edit tiba" hilang,kalo buat jj sering lambat gak mengikuti irama lagu nya	dulu bagus bangetsekarang sering nge lagterus kadang foto udah di edit tiba hilangkalo buat jj sering lambat gak mengikuti irama lagu nya
jagan di download!!apa apa harus pro dulu,buat video tugas malah jadi gagal,harga nya juga mahal buat di kantong pelajar.trus banyak iklan padahal dulu gak.sekarang juga ngelag banget gak jelas gak kek kayak dulu,mending gitu uninstall aja	jagan di downloadapa apa harus pro dulubuat video tugas malah jadi gagalharga nya juga mahal buat di kantong pelajartrus banyak iklan padahal dulu gaksekarang juga ngelag banget gak jelas gak kek kayak dulumending gitu uninstall aja
Capcut nggak sebagus dulu,pas mau gunain lagu malah munculnya nggak jelas.sering muncul iklan mana iklan nya berulang ulang iklan nya lama lagi.	capcut nggak sebagus dulupas mau gunain lagu malah munculnya nggak jelassering muncul iklan mana iklan nya berulang ulang iklan nya lama lagi

- **Tokenization**

  Tokenization is the process of breaking down a review text into smaller units called tokens. Tokens are usually individual words that will be analyzed further. For example, the sentence "CapCut sangat mudah digunakan" will be transformed into tokens ['capcut', 'sangat', 'mudah', 'digunakan'].

Table 3: Tokenization

Ulasan	Case Folding
"Awalnya CapCut sangat bagus dan mudah digunakan, tapi sekarang semakin mengecewakan. Terlalu banyak iklan yang muncul, setiap kali mau edit selalu terganggu. Selain itu, fitur-fitur penting seperti filter dan efek kini banyak yang dikunci di CapCut Pro. Dikit-dikit disuruh bayar untuk akses premium, padahal dulunya semua ini gratis. Aplikasi ini sudah tidak ramah pengguna lagi. "	['capcut', 'bagus', 'mudah', 'mengecewakan', 'iklan', 'muncul', 'kali', 'edit', 'terganggu', 'fiturfitur', 'filter', 'efek', 'dikunci', 'capcut', 'pro', 'dikitdikit', 'disuruh', 'bayar', 'akses', 'premium', 'dulunya', 'gratis', 'aplikasi', 'ramah', 'pengguna']
"CapCut adalah aplikasi editing video komplit tidak berbayar dan Anda butuhkan untuk membuat video berkualitas tinggi yang menakjubkan." Tapi sayang kenapa ada fitur watermark saat export video mana harus bayar lagi... Katanya gratis tapi kok berbayar GK sesuai di deskripsi	['capcut', 'aplikasi', 'editing', 'video', 'komplit', 'berbayar', 'butuhkan', 'video', 'berkualitas', 'menakjubkan', 'sayang', 'fitur', 'watermark', 'export', 'video', 'bayar', 'gratis', 'berbayar', 'gk', 'sesuai', 'deskripsi']
"CapCut sangat membantu saya dalam membuat video kreatif dengan mudah. Fitur teks dan efek videonya sangat keren dan intuitif. Namun, terkadang aplikasi terasa sedikit lambat saat mengedit video panjang. Secara keseluruhan, saya sangat puas dan merekomendasikan aplikasi ini."	['capcut', 'membantu', 'video', 'kreatif', 'mudah', 'fitur', 'teks', 'efek', 'videonya', 'keren', 'intuitif', 'terkadang', 'aplikasi', 'lambat', 'mengedit', 'video', 'puas', 'merekomendasikan', 'aplikasi']

- **Stopword Removal**

  At this stage, common words that do not have a significant impact on the analysis, such as "and," "which," or "in," are removed. This step is carried out to focus the analysis on more relevant keywords that better describe the sentiment of the review.

Table 4: Stopword Removal

Ulasan	Case Folding
Abis di update malah jadi banyak iklan Mau simpan video harus nonton iklan dulu , sebelum di update juga ga gitu Jadi ga jelas aplikasinya se habis di update ,baru buka cap cut langsung di sambut iklan mana lama banget nunggunya belum lagi kalo mau simpen video tolong di buat kaya dulu lagi biar gada iklan .	abis update iklan simpan video nonton iklan update ga gitu ga aplikasinya habis update buka cap cut langsung sambut iklan banget nunggunya kalo simpen video
Abis diupdate bukannya nambah stabil malah makin amburadul, ketika mau masuk yang muncul hanya layar putih, ditunggu makin lama ga berubah juga, iklan makin banyak, belum edit pun udah dikasih iklan.	abis diupdate nambah stabil amburadul masuk muncul layar putih ditunggu ga berubah iklan edit udah dikasih iklan
Abis update malah jadi aneh, masa setiap export hasil ny ada blank hitam beberapa detik. Durasi 15 detik jadi 18 detik. Thumb naik sudah di setting tp hasil akhir beda. Hadehhh 😭	abis update aneh export hasil ny blank hitam detik durasi detik detik thumb setting tp hasil beda hadehhh

- **Stemming**
  Stemming is the process of transforming each word into its base form. For example, the words "digunakan" (used), "menggunakan" (using), and "penggunaan" (usage) will be simplified to the root word "guna" (use). This process aims to help the model recognize relationships between words more effectively.

Table 5: Stemming

Ulasan	Case Folding
"Awalnya CapCut sangat bagus dan mudah digunakan,tapi sekarang semakin mengecewakan. Terlalu banyak iklan yang muncul, setiap kali mau edit selalu terganggu. Selain itu, fitur-fitur penting seperti filter dan efek kini banyak yang dikunci di CapCut Pro. Dikit-dikit disuruh bayar untuk akses premium, padahal dulunya semua ini gratis. Aplikasi ini sudah tidak ramah pengguna lagi. "	capcut bagus mudah kecewa iklan muncul kali edit ganggu fiturfitur filter efek kunci capcut pro dikitdikit suruh bayar akses premium dulunya gratis aplikasi ramah guna
Aelah, apaan update sekarang. Mau simpan video edit perlu nonton iklan. Banyak banget sekarang iklan yah. No Watermark pun harus premium. Haduh, kacau. Nih aplikasi lama lama di tinggal orang tau gk, kalo cuma mau untung mulu	aelah update simpan video edit nonton iklan banget iklan yah no watermark premium haduh kacau nih aplikasi tinggal orang tau gk kalo untung mulu
Aga gasuka soal nya terlalu banyak iklan kadang iklan nya ga selesai selesai untuk operator perbaiki lagi iklan nya kurangin kalo perlu gausah pake iklan	aga gasuka nya iklan kadang iklan nya ga selesai selesai operator baik iklan nya rangin kalo gausah pake iklan

3.1.3. Tranformation Data

After the preprocessing stage, the text data is converted into a numerical format using the TF-IDF (Term Frequency-Inverse Document Frequency) method. TF-IDF assigns a higher weight to words that frequently appear in a specific document but are rare across the entire dataset. This representation helps the model focus on unique words that are directly related to the user's sentiment.

Fig. 2: Positive Review

Fig. 3: Negative Review

3.1.4. Model Implementation

Before training the model, the dataset is split into training and testing data. This split is done using the train_test_split function from the scikit-learn library. Here are the steps:

Fig. 4: Data Training and Data Test

- **Data Splitting**
  The data is divided into two parts with an 80% proportion for training and 20% for testing, using the parameter test_size=0.20 and random_state=0 to ensure consistent results. The feature used is the 'content' column, while the label is the 'Label' column.
- **Data Transformation with TF-IDF**
  The text features in the training and testing data are converted into numerical representations using the TF-IDF Vectorizer from the scikit-learn library.
- **Data Dimensions**
  The data dimensions after splitting and transformation are as follows:
  X_train.shape: Displays the dimensions of the training feature data (685,).
  y_train.shape: Displays the dimensions of the training label data (685,).
  X_test.shape: Displays the dimensions of the testing feature data (172,).
  y_test.shape: Displays the dimensions of the testing label data (172,).

3.1.5. Model evaluation

The model's performance is evaluated using standard metrics such as accuracy, precision, recall, and F1-score. Accuracy indicates how often the model makes correct predictions, while precision and recall measure the model's ability to recognize specific sentiment categories. The F1-score, which is a combination of precision and recall, provides an overall view of the model's performance.

	precision	recall	f1-score	support
Negatif	0.90	1.00	0.95	152
Positif	1.00	0.24	0.38	21
accuracy			0.91	173
macro avg	0.95	0.62	0.67	173
weighted avg	0.92	0.91	0.88	173

Fig. 5: classification report

3.1.6. Interpretation

After evaluation, the sentiment analysis results are interpreted to provide insights into user views on the CapCut application. The majority of reviews show positive sentiment, reflecting satisfaction with features such as ease of use, attractive design, and video sharing capabilities. However, some negative reviews identify issues such as bugs, intrusive ads, and limitations in certain features. These findings offer a clear picture of aspects that users like and areas that need improvement in the CapCut app. The details are shown in the table below.

- **Accuracy**: Measures the percentage of correct predictions across the entire dataset. This metric provides an overall view of the model's ability to classify reviews into positive or negative sentiment.
- **Precision**: Indicates the accuracy of the model in identifying a specific category, such as positive sentiment. Precision is calculated as the ratio of correct predictions for that category to the total predictions for the same category.
- **Recall**: Measures the model's ability to identify all data that genuinely belongs to a specific category. Recall reflects the model's sensitivity to the category being analyzed.
- **F1-Score**: The harmonic mean between precision and recall. This metric is used to evaluate the balance between the two. The model's assessment is done using the test data to ensure the model performs optimally on new data and avoids overfitting. The evaluation process is also supported by cross-validation techniques to improve the accuracy of the results.
-

Table 6: Interpretation

	Precision	**Recall**	**F1-score**	**Suport**
Negatif	83%	100%	96%	141%
Positif	100%	22%	36%	30%
Accuracy			83%	171%
Marco Avg	91%	52%	49%	171%
Weighted Avg	86%	83%	76%	171%

3.2. Discusion

This sentiment analysis identifies the factors influencing users to either praise or criticize the CapCut app. Positive reviews generally praise the app for its ease of use, innovative editing features, and practical video-sharing capabilities. Negative reviews, however, highlight technical issues such as bugs, excessive ads, and the transition from free features to paid ones (CapCut Pro). A study by Sari et al. (2023) found that user reviews often correlate with app performance, with technical problems like bugs and intrusive ads being commonly mentioned in negative reviews. The model used in this analysis, with high accuracy, successfully distinguishes between positive and negative reviews and identifies the causes of user dissatisfaction, such as bugs and ads. The model achieved a perfect recall of 100%, detecting all negative reviews.The model's performance was evaluated using metrics like Confusion Matrix, Accuracy, Precision, and Recall. Key findings include:

Accuracy	=	$\frac{TP+TN}{TP+TN+FP+FN}$	$\frac{54+154}{41+154FP+FN} x100\% = 91.86\%$
Precision	=	$\frac{TP}{TP+FP}$	$\frac{41}{41+FP} x100\% = 91.67\%$
Recall	=	$\frac{TP}{TP+FN}$	$\frac{41}{41+0} x100\% = 100\%$

- **Accuracy (91.86%)**: The model correctly predicted 91.86% of the total reviews, reflecting strong overall performance.
- **Precision (91.67%)**: This indicates that 91.67% of the reviews predicted as negative by the model were accurate, with minimal false positives.
- **Recall (100%)**: The model detected all negative reviews without missing any, crucial for ensuring all criticisms are identified.
- **True Positive (41)**: The model accurately identified 41 negative reviews.
- **True Negative (154)**: The model correctly classified 154 non-negative reviews.

4.Conclusion

To develop an efficient and automated method for analyzing user sentiment in CapCut reviews on the Play Store, implementing effective machine learning algorithms like Naive Bayes is crucial. The process involves collecting user reviews through web scraping, preprocessing text to ensure clean and structured data, and utilizing word-weighting techniques such as TF-IDF for feature representation. This approach allows for automatic sentiment classification into positive or negative categories. The performance of the method can be evaluated using metrics like accuracy, precision, and recall to ensure accurate and efficient results.

Sentiment analysis reveals that users appreciate CapCut for its user-friendly interface, accessibility, and appealing visual effects. On the other hand, criticisms often focus on intrusive ads, bugs that disrupt the user experience, and the limited functionality caused by premium features. Understanding these patterns of appreciation and criticism provides valuable insights for developers to enhance standout features and address shortcomings, ensuring the app continues to evolve to meet user expectations.

References

1. E. W. Sholeha, S. Yunita, R. Hammad, V. C. Hardita, and K. Kaharuddin, "Analisis Sentimen Pada Agen Perjalanan Online Menggunakan Naïve Bayes dan K-Nearest Neighbor," *JTIM J. Teknol. Inf. dan Multimed.*, vol. 3, no. 4, pp. 203–208, 2022, doi: 10.35746/jtim.v3i4.178.
2. Syahril Dwi Prasetyo, Shofa Shofiah Hilabi, and Fitri Nurapriani, "Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN," *J. KomtekInfo*, vol. 10, pp. 1–7, 2023, doi: 10.35134/komtekinfo.v10i1.330.
3. P. Sari, E. Qomariah, M. Jazuli, and ..., "Pemanfaatan Aplikasi Android dalam Promosi Kain Sasirangan oleh Pengrajin di Bantaran Sungai Lulut sebagai Digital Startup di Masa Pandemi Covid-19," *Pros. Semin. …*, pp. 612–622, 2023.
4. V. F. Anindya *et al.*, "JIFSI: Jurnal Informatika dan Sistem Informasi Analisis Sentimen Ulasan Aplikasi Capcut Menggunakan Metode Naive Bayes," vol. 3, no. 1, pp. 24–33, 2023.
5. R. Fachrina and Z. M.Nawawi, "Pemanfaatan Digital Marketing (Shopee) Dalam Meningkatkan Penjualan Pada UMKM Di Marelan," *J. Ilm. Mhs. Perbank. Syariah*, vol. 2, no. 2, pp. 247–254, 2022, doi: 10.36908/jimpa.v2i2.75.
6. K. A. Nugraha, "Analisis Sentimen Berbasis Emoticon pada Komentar Instagram Bahasa Indonesia Menggunakan Naïve Bayes," *J. Tek. Inform. dan Sist. Inf.*, vol. 7, no. 3, pp. 715–721, 2021, doi: 10.28932/jutisi.v7i3.4094.
7. Merinda Lestandy, Abdurrahim Abdurrahim, and Lailis Syafa'ah, "Analisis Sentimen Tweet Vaksin COVID-19 Menggunakan Recurrent Neural Network dan Naïve Bayes," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 4, pp. 802–808, 2021, doi: 10.29207/resti.v5i4.3308.
8. D. Al Ghozi and F. N. Hasan, "Implementation of User Sentiment with Naïve Bayes Algorithm to Analyze LinkedIn Application Regarding Job Vacancies in the Play Store," *J. Media Inform. Budidarma*, vol. 8, no. 3, p. 1647, 2024, doi: 10.30865/mib.v8i3.7879.
9. Abdullah, D., Zarlis, M., Pardede, A. M. H., Anum, A., Suryani, R., Parwito, ... & Setiyadi, D. (2019, November). Expert System Diagnosing Disease of Honey Guava Using Bayes Method. In *Journal of Physics: Conference Series* (Vol. 1361, No. 1, p. 012054). IOP Publishing.
10. Lestari, S. A. M., Pardede, A. M., & Simanjuntak, M. (2024). Prediksi Disleksia pada Anak menggunakan Metode Naive Bayes. *Jurnal Kajian dan Penelitian Umum*, *2*(5), 37-51