Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 2, April 2024
https://doi.org/10.47709/cnahpc.v6i2.3990

Submitted : Mei 17, 2024
Accepted : Mei 30, 2024
Published : June 06, 2024

Analysis of Public Sentiment Towards The Tiktok Application Using The Naive Bayes Algorithm and Support Vector Machine

Ika Arofatul Hidayah1)*, Ririen Kusumawati2), Zainal Abidin3), M. Imamuddin4)
1)*2)3)4)Faculty of Science and Technology, Universitas Islam Negeri Maulana Malik Ibrahim Malang, East Java, Indonesia
1)* ikaarofatul@gmail.com, 2) ririen.kusumawati@ti.uin-malang.ac.id, 3) zainal@ti.uin-malang.ac.id,
4) imamudin@ti.uin-malang.ac.id

ABSTRACT

In the current digital era, social media applications such as TikTok have become an important aspect of people's lives. TikTok allows users to create and share short videos, making it a global phenomenon with millions of active users. However, this application has also been the subject of various responses and opinions from the public. This research aims to classify public sentiment towards the TikTok application based on comments on Playstore using the Naïve Bayes algorithm and Support Vector Machine (SVM). This research method involves collecting comment data from Playstore using scraping techniques, resulting in 5,000 review data. Data pre-processing stages include case folding, tokenization, normalization, stopword removal, stemming, and data labeling using a lexicon. The data that has been processed is then weighted using Term Frequency - Inverse Document Frequency (TF-IDF) before being classified using the Naïve Bayes and SVM algorithms. Algorithm performance evaluation is carried out using the Confusion Matrix to measure accuracy, precision and recall. The research results show that the SVM algorithm has higher accuracy (84%) compared to Naïve Bayes (79%). SVM also shows better precision and recall values in classifying positive and negative sentiment from user reviews. From the results of the tests that have been carried out, the SVM algorithm is more effective than Naïve Bayes in sentiment analysis of the TikTok application. This research provides insight into how public sentiment can be measured and analyzed, and underscores the importance of choosing the right algorithm for data sentiment analysis on social media platforms.

Keywords: Sentiment Analysis; Naïve Bayes; Support Vector Machine; TikTok

INTRODUCTION

TikTok is a social media platform that allows users to create and share short videos of short duration, usually ranging from 15 to 60 seconds (Indriyani, Fauzi, & Faisal, 2023). The app has become a global phenomenon with millions of active users worldwide. TikTok users can create various types of content, ranging from dance, lip-sync, comedy, tutorials, educational content, to the latest trends. TikTok offers a variety of creative features and editing tools that allow users to easily create engaging and entertaining videos (Isnan, Elwirehardja, & Pardamean, 2023). However, as with other social media platforms, TikTok is also the subject of various responses and opinions from the public.

Sentiment analysis is a Natural Language Processing (NLP) technique used to determine the emotional tone behind a text. It is an important tool that can help understand public opinion, emotions, and responses to a product, service, or idea (Fide, Suparti, & Sudarno, 2021). This analysis is often used to examine reviews, feedback, and social media to gain insight into public attitudes. With the growth of unstructured data on the internet, such as comments and posts on social media, sentiment analysis has become key to business strategy and social research, allowing organizations to respond more effectively to customer needs and wants.

TikTok has become a global phenomenon that revolutionizes the use of social media, making it a very interesting research topic. As a platform that attracts millions of active users, TikTok introduced an innovative short video format, which has changed the way content is created and consumed. Users' interactions with the platform reflect a shift in digital media consumption, particularly among the younger generation, who prefer content that is fast, accessible and highly interactive. TikTok is not only an entertainment platform but also a space for political, social, and cultural expression that allows users to share and influence opinions globally.

* Corresponding author

Submitted : Mei 17, 2024
Accepted : Mei 30, 2024
Published : June 06, 2024

An analysis of sentiment towards TikTok can provide insights into how social values and norms are shaped and reflected through modern digital media. The study of public sentiment towards TikTok can also reveal how the app influences issues such as generational identity, intercultural interactions, and the dynamics of globalization. The characteristics of comments on the TikTok app in the Playstore are particularly interesting to study as they reflect users' diverse perceptions of the app. These comments include aspects such as user satisfaction with new features, complaints about bugs or performance issues, and responses to the app's privacy and security policies. Analysis of these comments can provide insight into how changes or updates to the app affect user sentiment.

Based on the description above, this research tries to classify public sentiment towards the TikTok application based on comments found on Playstore using the naïve bayes algorithm and support vector machine. The results of this study are expected to determine the performance of the naïve bayes algorithm and support vector machine algorithm in classifying positive and negative and producing a comparison of the accuracy, precision and recall values of the data.

LITERATURE REVIEW

There are several studies related to this research, including: Sola Fide et al (2021) analyzed public sentiment towards TikTok in Indonesia using SVM with Radial Basis Function kernel, combining association methods. The research process involved scrapping data from Google Play, preprocessing (case folding, cleaning, normalization), sentiment scoring, feature selection, TF-IDF weighting, and generation and classification of training and test data. SVM yielded an accuracy of 90.62% and kappa of 81.24%, showing high effectiveness in identifying positive sentiments related to content and negative ones related to technical issues such as account registration and blocking.

Friska Aditia Indriyani et al (2023) conducted sentiment analysis on the TikTok app by categorizing reviews into positive and negative, using Naïve Bayes and Support Vector Machine algorithms. The data, consisting of 2000 relevant comments, was collected through web scraping and processed through pre-processing including Case Folding, Tokenization, Stop Word Removal, and Stemming, with weighting using TF-IDF. Classification was performed with both algorithms and their effectiveness was measured using Confusion Matrix. Results showed that SVM with 84% accuracy was more effective than Naïve Bayes, which achieved 79% accuracy.

Junda Alfiah Zulqornain et al (2021) analyzed public sentiment towards the TikTok application, which often displays vulgar content without age restrictions, using the Naïve Bayes and Categorical Proportional Difference algorithms. The aim of the study was to help parents choose appropriate apps for their children. Data from Google PlayStore included 1000 reviews with a score of 1-5, classified into positive and negative. Results showed 73% accuracy, 75% precision, 93% recall, and 82% f-measure.

Dian Ardiansyah et al (2023) analyzed sentiment towards the TikTok application using the K-Nearest Neighbor and Support Vector Machine (SVM) algorithms, which were optimized with Particle Swarm Optimization (PSO). They compared the performance of these two algorithms in identifying positive and negative sentiments based on review data from Playstore. The research process includes data scrapping, manual labeling, preprocessing, and evaluation using Cross Validation. The results show that SVM with PSO has the highest accuracy of 88.20% and AUC 0.91, more effective than K-Nearest Neighbor which has a maximum accuracy of 83.40% and AUC 0.903.

Winda Yulita et al (2021) analyzed the sentiment of Indonesian people about the COVID-19 vaccine using the Naïve Bayes Classifier algorithm. In the context of a pandemic that caused lockdowns and significant reactions on social media to the first vaccine announcement, this study processed and analyzed 3780 tweets using techniques such as cleansing, tokenization, and TF-IDF. Evaluation was performed using Confusion Matrix, showing that 60.3% of tweets were positive, 34.4% neutral, and 5.4% negative towards vaccination, with analysis accuracy reaching 93%.

Hindun Habibatul Mubaroroh et al (2022) analyzed public sentiment towards the Ruangguru platform using Naïve Bayes Classifier enriched with Levenshtein Distance word normalization. Out of 1500 reviews, 1484 spam-free ones were used as research data. Preprocessing methods included Levenshtein Distance, case folding, cleansing, stopwords removal, stemming, tokenizing, and TF-IDF. Evaluation using 10-fold cross validation showed that the Naïve Bayes algorithm achieved an average accuracy of 88.20%, with a peak of 94% at the eighth fold, proving its effectiveness in classifying Ruangguru reviews.

METHOD

This research requires the correct procedure so that this research can run effectively. The following flow of this research can be seen in Figure 1:

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 2, April 2024
https://doi.org/10.47709/cnahpc.v6i2.3990

Submitted : Mei 17, 2024
Accepted : Mei 30, 2024
Published : June 06, 2024

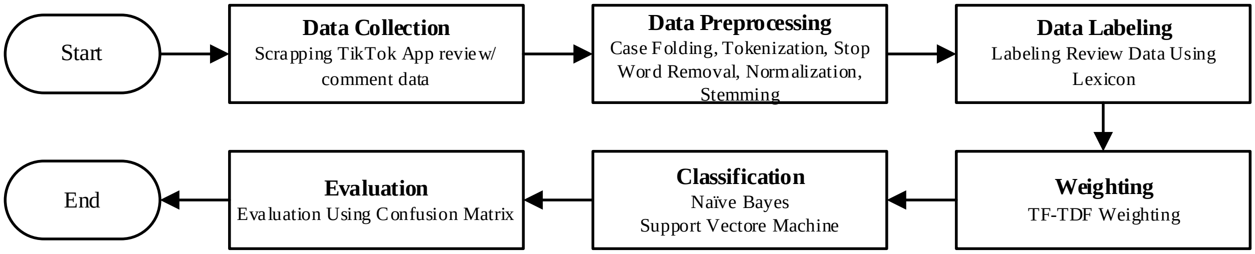
Start
Data Collection
Scrapping TikTok App review/ comment data
Data Preprocessing
Case Folding, Tokenization, Stop Word Removal, Normalization, Stemming
Data Labeling
Labeling Review Data Using Lexicon
Weighting
TF-TDF Weighting
Classification
Naïve Bayes
Support Vectore Machine
Evaluation
Evaluation Using Confusion Matrix
End

Fig.1 Flow of Research

The data used in this study uses data on reviews / comments from Indonesian people on the Tik-Tok application on Google PlayStore. Data collection is done by scrapping Tik-Tok comment data using Google Collaboratory with the determination of only scrapping with id-Indonesia (Indonesian comments only). The amount of data collected is 313,187 data from June 28, 2023 to December 31, 2023, but the data used in this study is 5,000 data.

Preprocessing is the process of converting bad data into good data to obtain an authentic source of information that will be processed further. At this stage, several stages are carried out including cleaning, repairing and combining data. This research will preprocess the data by breaking the sentence into words (Tokenization), then the word repair process is carried out updating the typo word to be correct (Spell Checker) and repairing words that are not standard to be standard (Normalization), and removing words that have many occurrences / conjunctions (Stopword Removal) and removing affix words from words that have affixes (Stemming).

It functions to identify a word in a sentence that has a positive, negative or neutral nature by calculating the polarity value. The way to identify the word to know the sentiment of the word, can use the Indonesian language lexicon dataset. The first step in this process is the development or selection of an appropriate lexicon. Once the lexicon is prepared, the labeling process begins with the analysis of the text to be labeled. The text is broken down into smaller units, usually words, which are then matched with entries in the lexicon. Each word in the text is assigned a score based on its associated sentiment value in the lexicon. For example, the word “happy” may have a positive value, while the word “sad” has a negative value. This process allows the researcher to automatically collect scores from all the words in the text and aggregate the scores to get an overall sentiment representation of the text.

Term Frequency - Inverse Document Frequency word weighting which aims to calculate the weight value of each word in each document. At this stage it is divided into 2 processes, namely Term Frequency and Inverse Document Frequency. TF (Term Frequency) calculates the number of occurrences of each word in the document and with the most occurrences of the word, the value of the word is the greatest. IDF (Inverse Document Frequency) calculates the number of documents in each word that rarely appears in a document which is considered the greatest value. If the word has many occurrences in the document then the value is small.

Testing is carried out using the Naïve Bayes and Support Vector Machine algorithms in the process of classifying public sentiment towards the TikTok application. The 80:20 data ratio used is 80% as training data and 20% as testing data. To determine the performance of the Naive Bayes and Support Vector Machine algorithms, this research uses Confusion Matrix. Confusion Matrix uses a matrix table to show the classification results of testing data based on training data. The following is a Confusion Matrix image that has four different combinations of predicted and actual values.

* Corresponding author

Submitted : Mei 17, 2024
Accepted : Mei 30, 2024
Published : June 06, 2024

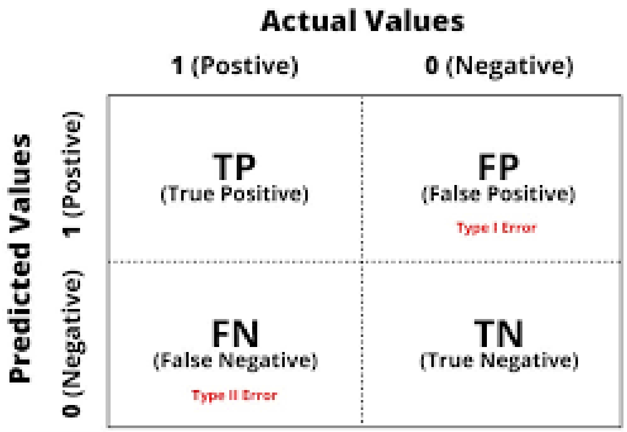

Fig. 2 Confusion Matrix

Description. TP is positive data that is predicted correctly, TN is negative data that is correctly predicted, FP is negative data but predicted as positive data, FN is positive data but predicted as negative data.

RESULT

Data Collection

The TikTok application review or comment data used in this study is 5,000 data. The following part of the TikTok review/comment data that has been collected can be seen in table 1:

Table 1
TikTok App Review/Comment Data

No	Komentar	Skor	Tanggal
1	baguuzzz bgtttz aks sukaaaak😜😝❤️‍🔥❤️‍🔥	5	2023-12-31 23:59:24
2	Aplikasi gak jelas, padahal enggak pernah ngirim/ngetik aneh aneh malah ditangguhkan nungguin lama tetep gak bisa, gak bisa ngirim pesan, enggak bisa terima pesan tolong ya aplikasi nya di perbaiki lagi	1	2023-12-31 23:58:40
3	Skrang Tiktok Terlalu sensitif biar kata tidak mengandung kata Negatif... Biar pun bahasa daerah yang kata membangun Semangat... Sedikit-Sediki di hapus.. Biar pun di Banding" tetap aja di hapus....Tolong di perhatikan dan belajar bahasa daerah	1	2023-12-31 23:58:33
4	Keren	4	2023-12-31 23:52:35
5	Terlalu banyak pelanggaran yg di berikan Padahal vidio buat sendiri tapi kena pelanggaran sampai 3 point Tolong di perbaiki sistem nya	1	2023-12-31 23:52:09

Data Preprocessing

The first stage before the data classifier process is preprocessing. Preprocessing stages carried out include case folding, tokenization, stopword removal and stemming.

Table 2

* Corresponding author

Submitted : Mei 17, 2024
Accepted : Mei 30, 2024
Published : June 06, 2024

Case Folding Result

No	Comment	Case Folding
1	baguuzzz bgtttz aks sukaaaak😜 😝 ♡ ♡	baguuzzz bgtttz aks sukaaaak
2	Aplikasi gak jelas, padahal enggak pernah ngirim/ngetik aneh aneh malah ditangguhkan nungguin lama tetep gak bisa, gak bisa ngirim pesan, enggak bisa terima pesan tolong ya aplikasi nya di perbaiki lagi	aplikasi gak jelas, padahal enggak pernah ngirim/ngetik aneh aneh malah ditangguhkan nungguin lama tetep gak bisa, gak bisa ngirim pesan, enggak bisa terima pesan tolong ya aplikasi nya di perbaiki lagi
3	Skrang Tiktok Terlalu sensitif biar kata tidak mengandung kata Negatif... Biar pun bahasa daerah yang kata membangun Semangat... Sedikit-Sediki di hapus.. Biar pun di Banding" tetap aja di hapus....Tolong di perhatikan dan belajar bahasa daerah	skrang tiktok terlalu sensitif biar kata tidak mengandung kata negatif... biar pun bahasa daerah yang kata membangun semangat... sedikit-sediki di hapus.. biar pun di banding" tetap aja di hapus....tolong di perhatikan dan belajar bahasa daerah
4	Keren	keren
5	Terlalu banyak pelanggaran yg di berikan Padahal vidio buat sendiri tapi kena pelanggaran sampai 3 point Tolong di perbaiki sistem nya	terlalu banyak pelanggaran yg di berikan padahal vidio buat sendiri tapi kena pelanggaran sampai 3 point tolong di perbaiki sistem nya

Table 2 is the result of the case folding process in this process aims to convert words into lowercase letters, and delete all sentences other than text such as punctuation marks, symbols, emojis and others, also deleting numbers in a sentence in the document. The next stage of pre-processing is tokenizing, in Table 3 it can be seen that this process breaks a sentence into a token so that it can be processed at the next stage.

Table 3
Tokenization Result

No	Case Folding	Tokenization
1	baguuzzz bgtttz aks sukaaaak	['baguuzzz', 'bgtttz', 'aks', 'sukaaaak']
2	aplikasi gak jelas, padahal enggak pernah ngirim/ngetik aneh aneh malah ditangguhkan nungguin lama tetep gak bisa, gak bisa ngirim pesan, enggak bisa terima pesan tolong ya aplikasi nya di perbaiki lagi	['aplikasi', 'gak', 'jelas', 'padahal', 'enggak', 'pernah', 'ngirim', 'ngetik', 'aneh', 'aneh', 'malah', 'ditangguhkan', 'nungguin', 'lama', 'tetep', 'gak', 'bisa', 'gak', 'bisa', 'ngirim', 'pesan', 'enggak', 'bisa', 'terima', 'pesan', 'tolong', 'ya', 'aplikasi', 'nya', 'di', 'perbaiki', 'lagi']
3	skrang tiktok terlalu sensitif biar kata tidak mengandung kata negatif... biar pun bahasa daerah yang kata membangun semangat... sedikit-sediki di hapus.. biar pun di banding" tetap aja di hapus....tolong di perhatikan dan belajar bahasa daerah	['skrang', 'tiktok', 'terlalu', 'sensitif', 'biar', 'kata', 'tidak', 'mengandung', 'kata', 'negatif', 'biar', 'pun', 'bahasa', 'daerah', 'yang', 'kata', 'membangun', 'semangat', 'sedikit', 'sediki', 'di', 'hapus', 'biar', 'pun', 'di', 'banding', 'tetap', 'aja', 'di', 'hapus', 'tolong', 'di', 'perhatikan', 'dan', 'belajar', 'bahasa', 'daerah']
4	keren	[keren]
5	terlalu banyak pelanggaran yg di berikan padahal vidio buat sendiri tapi kena pelanggaran sampai 3 point tolong di perbaiki sistem nya	['terlalu', 'banyak', 'pelanggaran', 'yg', 'di', 'berikan', 'padahal', 'vidio', 'buat', 'sendiri', 'tapi', 'kena', 'pelanggaran', 'sampai', '3', 'point', 'tolong', 'di', 'perbaiki', 'sistem', 'nya']

Table 4 is the result of the normalization. Normalization in this study was carried out by converting non-standard

Submitted : Mei 17, 2024
Accepted : Mei 30, 2024
Published : June 06, 2024

words into standard words in accordance with the Big Indonesian Dictionary (KBBI) by matching with the dataset that was already owned (informal sentences - formal sentences).

Table 4
Normalization Result

No	Tokenization	Normalization
1	['baguuzzz', 'bgtttz', 'aks', 'sukaaaak']	['bagus', 'banget', 'aku', 'suka']
2	['aplikasi', 'gak', 'jelas', 'padahal', 'enggak', 'pernah', 'ngirim', 'ngetik', 'aneh', 'aneh', 'malah', 'ditangguhkan', 'nungguin', 'lama', 'tetep', 'gak', 'bisa', 'gak', 'bisa', 'ngirim', 'pesan', 'enggak', 'bisa', 'terima', 'pesan', 'tolong', 'ya', 'aplikasi', 'nya', 'di', 'perbaiki', 'lagi']	['aplikasi', 'tidak', 'jelas', 'padahal', 'tidak', 'pernah', 'mengirim', 'mengetik', 'aneh', 'aneh', 'malah', 'ditangguhkan', 'menunggu', 'lama', 'tetap', 'tidak', 'bisa', 'tidak', 'bisa', 'mengirim', 'pesan', 'tidak', 'bisa', 'terima', 'pesan', 'tolong', 'ya', 'aplikasi', 'nya', 'di', 'perbaiki', 'lagi']
3	['skrang', 'tiktok', 'terlalu', 'sensitif', 'biar', 'kata', 'tidak', 'mengandung', 'kata', 'negatif', 'biar', 'pun', 'bahasa', 'daerah', 'yang', 'kata', 'membangun', 'semangat', 'sedikit', 'sediki', 'di', 'hapus', 'biar', 'pun', 'di', 'banding', 'tetap', 'aja', 'di', 'hapus', 'tolong', 'di', 'perhatikan', 'dan', 'belajar', 'bahasa', 'daerah']	['skrang', 'tiktok', 'terlalu', 'sensitif', 'biar', 'kata', 'tidak', 'mengandung', 'kata', 'negatif', 'biar', 'pun', 'bahasa', 'daerah', 'yang', 'kata', 'membangun', 'semangat', 'sedikit', 'sediki', 'di', 'hapus', 'biar', 'pun', 'di', 'banding', 'tetap', 'aja', 'di', 'hapus', 'tolong', 'di', 'perhatikan', 'dan', 'belajar', 'bahasa', 'daerah']
4	[keren]	[keren]
5	['terlalu', 'banyak', 'pelanggaran', 'yg', 'di', 'berikan', 'padahal', 'vidio', 'buat', 'sendiri', 'tapi', 'kena', 'pelanggaran', 'sampai', '3', 'point', 'tolong', 'di', 'perbaiki', 'sistem', 'nya']	['terlalu', 'banyak', 'pelanggaran', 'yang', 'di', 'berikan', 'padahal', 'video', 'buat', 'sendiri', 'tapi', 'kena', 'pelanggaran', 'sampai', '3', 'point', 'tolong', 'di', 'perbaiki', 'sistem', 'nya']

Table 5 is the result of the Stopword Removal process, at this stage it aims to take important words and discard words that are less important and have no meaning.

Table 5
Stopword Removal Result

No	Normalization	Stopword Removal
1	['bagus', 'banget', 'aku', 'suka']	['bagus', 'banget', 'suka']
2	['aplikasi', 'tidak', 'jelas', 'padahal', 'tidak', 'pernah', 'mengirim', 'mengetik', 'aneh', 'aneh', 'malah', 'ditangguhkan', 'menunggu', 'lama', 'tetap', 'tidak', 'bisa', 'tidak', 'bisa', 'mengirim', 'pesan', 'tidak', 'bisa', 'terima', 'pesan', 'tolong', 'ya', 'aplikasi', 'nya', 'di', 'perbaiki', 'lagi']	['aplikasi', 'tidak', 'jelas', 'padahal', 'tidak', 'pernah', 'mengirim', 'mengetik', 'aneh', 'aneh', 'malah', 'ditangguhkan', 'menunggu', 'lama', 'tetap', 'tidak', 'bisa', 'tidak', 'bisa', 'mengirim', 'pesan', 'tidak', 'bisa', 'terima', 'pesan', 'tolong', 'aplikasi', 'perbaiki', 'lagi']
3	['skrang', 'tiktok', 'terlalu', 'sensitif', 'biar', 'kata', 'tidak', 'mengandung', 'kata', 'negatif', 'biar', 'pun', 'bahasa', 'daerah', 'yang', 'kata', 'membangun', 'semangat', 'sedikit', 'sediki', 'di', 'hapus', 'biar', 'pun', 'di', 'banding', 'tetap', 'aja', 'di', 'hapus', 'tolong', 'di', 'perhatikan', 'dan', 'belajar', 'bahasa', 'daerah']	['sekarang', 'tiktok', 'terlalu', 'sensitif', 'kata', 'tidak', 'mengandung', 'kata', 'negatif', 'bahasa', 'daerah', 'kata', 'membangun', 'semangat', 'sedikit', 'sedikit', 'hapus', 'hapus', 'tolong', 'perhatikan', 'belajar', 'bahasa', 'daerah']
4	[keren]	[keren]
5	['terlalu', 'banyak', 'pelanggaran', 'yang', 'di', 'berikan', 'padahal', 'video', 'buat', 'sendiri', 'tapi', 'kena', 'pelanggaran', 'sampai', '3', 'point', 'tolong',	['terlalu', 'banyak', 'pelanggaran', 'berikan', 'padahal', 'video', 'buat', 'sendiri', 'kena', 'pelanggaran', 'sampai', '3', 'point', 'tolong', 'perbaiki', 'sistem']

No	Normalization	Stopword Removal
	'di', 'perbaiki', 'sistem', 'nya']	

Table 6 is the result of the stemming process at this stage of the process of combining basic words that have been filtered, in this process also sentences that were originally tokens are combined into ordinary sentences using the sastrawi library.

Table 6
Stemming Result

No	Stopword Removal	Stemming
1	['bagus', 'banget', 'suka']	['bagus', 'banget', 'suka']
2	['aplikasi', 'tidak', 'jelas', 'padahal', 'tidak', 'pernah', 'mengirim', 'mengetik', 'aneh', 'aneh', 'malah', 'ditangguhkan', 'menunggu', 'lama', 'tetap', 'tidak', 'bisa', 'tidak', 'bisa', 'mengirim', 'pesan', 'tidak', 'bisa', 'terima', 'pesan', 'tolong', 'aplikasi', 'perbaiki', 'lagi']	['aplikasi', 'tidak', 'jelas', 'padahal', 'tidak', 'pernah', 'kirim', 'tik', 'aneh', 'aneh', 'malah', 'tangguh', 'tunggu', 'lama', 'tetap', 'tidak', 'bisa', 'tidak', 'bisa', 'kirim', 'pesan', 'tidak', 'bisa', 'terima', 'pesan', 'tolong', 'aplikasi', 'baik', 'lagi']
3	['sekarang', 'tiktok', 'terlalu', 'sensitif', 'kata', 'tidak', 'mengandung', 'kata', 'negatif', 'bahasa', 'daerah', 'kata', 'membangun', 'semangat', 'sedikit', 'sedikit', 'hapus', 'hapus', 'tolong', 'perhatikan', 'belajar', 'bahasa', 'daerah']	['sekarang', 'tiktok', 'terlalu', 'sensitif', 'kata', 'tidak', 'mengandung', 'kata', 'negatif', 'bahasa', 'daerah', 'kata', 'bangun', 'semangat', 'sedikit', 'sedikit', 'hapus', 'hapus', 'tolong', 'perhatikan', 'belajar', 'bahasa', 'daerah']
4	[keren]	[keren]
5	['terlalu', 'banyak', 'pelanggaran', 'berikan', 'padahal', 'video', 'buat', 'sendiri', 'kena', 'pelanggaran', 'sampai', '3', 'point', 'tolong', 'perbaiki', 'sistem']	['terlalu', 'banyak', 'langgar', 'beri', 'padahal', 'video', 'buat', 'diri', 'kena', 'langgar', 'sampai', '3', 'point', 'tolong', 'baik', 'sistem']

Lexicon and Data Labeling

After the preprocessing stage to stemming, the next step is to match the words on the prepared Indonesian lexicon dataset to produce a sentence polarity value that will be used for determining positive and negative sentiments. The results of the Lexicon and Labeling process are shown in Figure 3.

GAMBAR 3

The comparison of the number of datasets after preprocessing and lexicon is 2579 negative sentiments (51.58%) and 2421 positive sentiments (48.42%) with a total of 5,000 data. The percentage of the data can be seen in Figure 4.

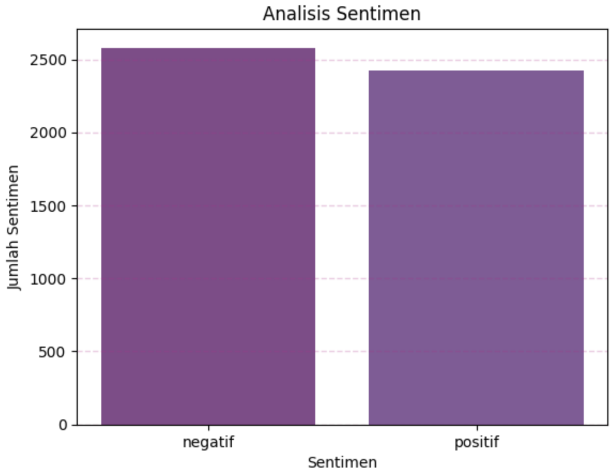

Fig. 3 Dataset Comparison Percentage

Weighting

Submitted : Mei 17, 2024
Accepted : Mei 30, 2024
Published : June 06, 2024

After performing the lexicon and data labeling stages, the next step is to calculate the weight of each term or word based on the frequency of occurrence of the term in the document using the TF-IDF method. TF-IDF is a numerical statistic that can show keywords with certain words. Apart from that, TF-IDF is also known to be efficient, simple and accurate. The TF-IDF data weighting formula is as follows:

$Wtt \times d = tft, d \times idft = tft, d \times \log N/dft.$

Description: (Wt,d) is TF-IDF Weight, (tft,d) is Number of word frequencies ($idft$) is Number of inverse document frequencies per word (dft) is Number of document frequencies per word, (N) is Total number of documents.

Data Split

The TikTok application comment/review text data classification process using Naive Bayes and Support Vector Machine is divided into 2 parts, namely training data and testing data. The data used is data that has been preprocessed and labeled. The dataset comparison used uses a ratio of 80% for training data and 20% for testing data. The comparison of the amount of data used can be seen in the following table 7.

Table 7
Division of Total Data

Label	Quantity	Training 80%	Testing 20%
Negatif	2579	2060	519
Positif	2421	1940	481
Total	**5000**	**4000**	**1000**

Classification Naïve Bayes

From the results of the experiments conducted, the Naive Bayes algorithm obtained quite good accuracy results in the process of classifying the sentiment of TikTok application reviews or comments with an accuracy rate of 79%. This is evidenced by the evaluation results of the Naive Bayes algorithm model testing process using 1000 testing data. Evaluation is done using Confusion Matrix to determine the performance of the algorithm. The results of the tests carried out can be seen in Figure 5.2:

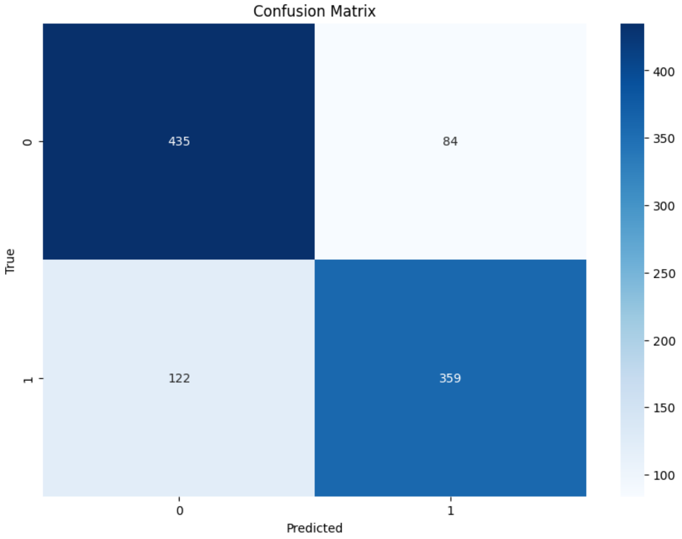

Fig. 4 Confusion Matrix of Naive Bayes Algorithm

Based on the confusion matrix above, it can be seen that the negative review data that is predicted to be negative (true) is 435 and the positive predicted (false) is 84. While the predicted positive review data is negative (wrong) as many as 122 and the predicted positive (correct) is 359. The precission, recall, and f1-score can be seen in table 8.

Table 8

* Corresponding author

Submitted : Mei 17, 2024
Accepted : Mei 30, 2024
Published : June 06, 2024

Naive Bayes Algorithm Test Results

Class	Precission	Recall	F1-Score
Negatif	0.78 (78%)	0.84 (84%)	0.81 (81%)
Positif	0.81 (81%)	0.75 (75%)	0.78 (78%)
Accuracy		**0.79 (79%)**	

Classification Support Vector Machine

From the results of the experiments conducted, the Support Vector Machine algorithm obtained very good accuracy results in the process of classifying sentiment reviews or comments on the TikTok application with an accuracy rate of 82%. This is evidenced by the evaluation results of the Support Vector Machine algorithm model testing process using 1000 testing data. Evaluation is done using Confusion Matrix to determine the performance of the algorithm. The results of the tests carried out can be seen in Figure 6.2:

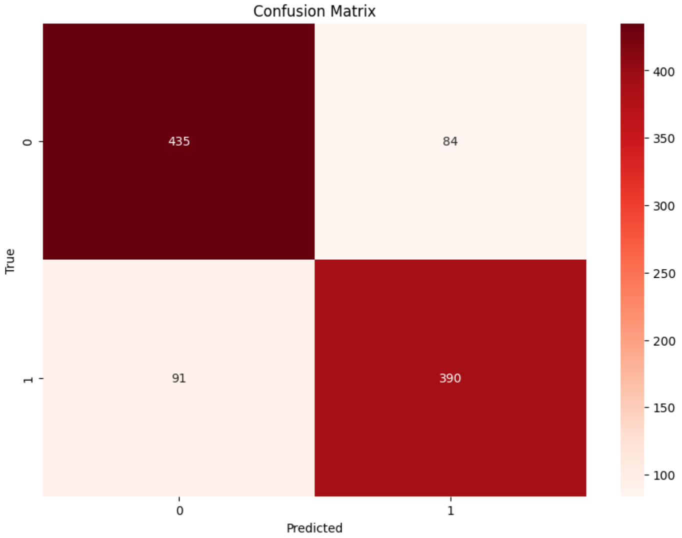

Fig. 5 Confusion Matrix of Support Vector Machine Algorithm

Based on the confusion matrix above, it can be seen that the negative review data that is predicted to be negative (true) is 435 and the positive predicted (false) is 84. While the predicted positive review data is negative (wrong) as many as 91 and the predicted positive (correct) is 390. The precission, recall, and f1-score can be seen in table 9.

Table 9
Support Vector Machine Algorithm Test Results

Class	Precission	Recall	F1-Score
Negatif	0.83 (83%)	0.84 (84%)	0.83 (83%)
Positif	0.82 (82%)	0.81 (81%)	0.82 (82%)
Accuracy		**0.82 (82%)**	

DISCUSSIONS

In an exploration of classification algorithms for managing sentiment data from TikTok app reviews, this research compares two popular methods: Naive Bayes and Support Vector Machine (SVM). Both algorithms are applied to the same dataset, which consists of 5000 data divided by a ratio of 80% for training and 20% for testing, to ensure both models are tested under the same conditions.

Naive Bayes, which is well-known for its efficiency in probability-based machine learning, showed adequate capability in sentiment classification by achieving 79% accuracy. Further analysis using confusion matrix shows that the model tends to identify negative reviews better than positive ones, with precision and recall of 78% and 84% for negative reviews, respectively. Although effective, this model has some limitations, such as the assumption of feature

* Corresponding author

Submitted : Mei 17, 2024
Accepted : Mei 30, 2024
Published : June 06, 2024

independence which may not always be appropriate in real sentiment analysis. In contrast, SVM, which uses a geometric approach to maximize the margin between data classes, managed to achieve a higher accuracy rate at 82%. The superiority of SVM in this study can be seen from its better ability to distinguish between positive and negative comments by using the optimal hyperplane. SVM showed consistent precision and recall above 80% for both classes, indicating a more balanced performance compared to Naive Bayes. The comparison of these two algorithmic models is as shown in Table 10.

Table 10
Comparison Table of Naive Bayes and SVM Algorithms

Criteria	Naive Bayes	SVM
Akurasi	79%	82%
Precision (Negatif)	78%	83%
Precision (Positif)	81%	82%
Recall (Negatif)	84%	84%
Recall (Positif)	75%	81%
F1-Score (Negatif)	81%	83%
F1-Score (Positif)	78%	82%

The comparison of these two models illustrates how algorithm selection can significantly impact sentiment classification results. While Naive Bayes is fast and efficient for large datasets, SVM offers higher accuracy with the ability to handle more complex and overlapped data. The decision between these two methods should be based on the specific needs of the application, where SVM may be more suitable for applications that require a high degree of discrimination between different sentiments. Both models, Naive Bayes and SVM, implement different approaches to the same problem, namely sentiment classification of TikTok app reviews. From the analysis conducted, Naive Bayes offers advantages in terms of speed and simplicity of implementation. The model is ideal for scenarios where processing speed is a critical factor, and available computing resources are limited. Although it has an independence assumption that could be a limitation in certain use cases, the technique is still relevant for datasets that have many features that are statistically independent.

On the other hand, SVM stands out for its ability to create strong decision boundaries, especially in cases where the data classes are not easily separated linearly. The higher accuracy in the SVM model suggests that this technique is able to overcome the weaknesses of Naive Bayes, especially in handling correlated features. Although it requires more computational resources and time, its effectiveness in classifying reviews more accurately makes it a better choice for applications that require a more precise level of prediction.

In conclusion, the choice between Naive Bayes and SVM should consider both the characteristics of the data at hand and the operational needs of the application using the model. In the context of sentiment analysis where nuance and context of words are important, SVM may offer a more robust approach. However, for implementations that require a quick response or where resources are limited, Naive Bayes can still be a very effective choice. As a next step, further testing with parameter variations and pre-modeling data processing techniques may provide deeper insights into how to improve the performance of these two models in various usage scenarios.

CONCLUSION

This research discusses in depth the performance of two classification algorithms, Naive Bayes and Support Vector Machine (SVM) in analyzing sentiment towards the TikTok application based on reviews/comments on Playstore. From this study, it was found that SVM showed better performance with 84% accuracy, while Naive Bayes achieved 79% accuracy. This shows that SVM is more effective in classifying positive and negative sentiments on the dataset used. This research provides insight into how these two algorithms process and categorize large and varied text data, which is important for the understanding of sentiment analysis in the digital age.

REFERENCES

Ardiansyah, D., Saepudin, A., Aryanti, R., Fitriani, E., & Royadi. (2023). Analisis Sentimen Review Pada Aplikasi Media Sosial Tiktok Menggunakan Algoritma K-NN dan SVM Berbasis PSO. *Jurnal Informatika Kaputama*

Submitted : Mei 17, 2024
Accepted : Mei 30, 2024
Published : June 06, 2024

(JIK), *7*(2), 233–241.

Fide, S., Suparti, & Sudarno. (2021). Analisis Sentimen Ulasan Aplikasi TikTok di Google Play Menggunakan Metode Support Vector Machine (SVM) dan Asosiasi. *Jurnal Gaussian*, *10*(3), 346–358. Retrieved from https://ejournal3.undip.ac.id/index.php/gaussian/

Hananto, B. K., Pinandito, A., & Kharisma, A. P. (2018). Penerapan Maximum TF-IDF Normalization Terhadap Metode KNN Untuk Klasifikasi Dataset Multiclass Panichella Pada Review Aplikasi Mobile. *Jurnal Pengembangan Teknologii Informasi Dan Ilmu Komputer*, *2*(12), 6812–6823. Retrieved from http://j-ptiik.ub.ac.id

Indriyani, F. A., Fauzi, A., & Faisal, S. (2023). Analisis Sentimen Aplikasi TikTok Menggunakan Algoritma Naive Bayes dan Support Vector Machine. *TEKNOSAINS : Jurnal Sains, Teknologi Dan Informatika*, *10*(2), 176–184. https://doi.org/10.37373/tekno.v10i2.419

Irfani, F. F., Triyanto, M., Hartanto, A. D., & Kusnawi. (2020). Analisis Sentimen Review Aplikasi Ruangguru Menggunakan Algoritma Support Vector Machine. *Jurnal Bisnis, Manajemen Dan Informatika*, *16*(3), 258–266.

Isnan, M., Elwirehardja, G. N., & Pardamean, B. (2023). Sentiment Analysis for TikTok Review Using VADER Sentiment and SVM Model. *Procedia Computer Science*, *227*, 168–175. https://doi.org/10.1016/j.procs.2023.10.514

Mubaroroh, H. H., Yasin, H., & Rusgiyono, A. (2022). Analisis Sentimen Data Ulasan Aplikasi Ruangguru Pada Situs Google Play Menggunakan Algoritma Naïve Bayes Classifier Dengan Normalisasi Kata Levenshtein Distance. *Jurnal Gaussian*, *11*(2), 248–257. Retrieved from https://ejournal3.undip.ac.id/index.php/gaussian/

Novitasari, I., Kurniawan, T. B., Dewi, D. A., & Misinem. (2022). Analisis Sentimen Masyarakat Terhadap Tweet Ruang Guru Menggunakan Algoritma Naive Bayes Classifier (NBC). *Jurnal Mantik*, *6*(3), 2685–4236.

Pawar, A. B., Jawale, M. A., & Kyatanavar, D. N. (2016). Fundamentals of Sentiment Analysis: Concepts and Methodology. In *Studies in Computational Intelligence* (Vol. 639, pp. 25–48). Springer Verlag. https://doi.org/10.1007/978-3-319-30319-2_2

Roldós, I. (2020, March 2). Go-to Guide for Text Classification with Machine Learning.

Samsir, Ambiyar, Verawardina, U., Edi, F., & Watrianthos, R. (2021). Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naive Bayes. *Jurnal Media Informatika Budidarma*, *5*(1), 157–163. https://doi.org/10.30865/mib.v5i1.2604

Wenando, F. A., Hayami, R., & Anggrawan, A. J. (2020). Analisis Sentimen Pada Pemerintahan Terpilih Pada Pilpres 2019 di Twitter Menggunakan Algoritman Naive Bayes. *Jurnal Teknologi Dan Sistem Informasi*, *7*(1), 101–106. https://doi.org/10.33330/jurteksi.v7i1.851

Worth, D. (2010). Introduction to Modern Information Retrieval, 3rd Edition. *Australian Academic and Research Libraries*, *41*(4), 305–306. https://doi.org/10.1080/00048623.2010.10721488

Yulita, W., Nugroho, E. D., & Algifari, M. H. (2021). Analisis Sentimen Terhadap Opini Masyarakat Tentang Vaksin Covid-19 Menggunakan Algoritma Naïve Bayes Classifier. *JDMSI*, *2*(2), 1–9.

Zulqornain, J. A., Indriati, & Adikara, P. P. (2021). Analisis Sentimen Tanggapan Masyarakat Aplikasi Tiktok Menggunakan Metode Naïve Bayes dan Categorial Propotional Difference (CPD). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer* , *5*(7), 2886–2890. Retrieved from http://j-ptiik.ub.ac.id