Jurnal Varian
Vol. 6, No. 1, Oktober 2022, Hal. 89–96
e-ISSN: 2581-2017

Sentiment Analysis Using Naive Bayes with Lexicon-Based Feature on TikTok Application

Siswanto[1], Zakiyah Marah[2], Alfiyah Salsa Dila Sabir[3], Taufik Hidayat[4], Fadilah Amirul Adhel[5], Waode Sitti Amni[6]

[1,3,4,5,6]Department of Statistics, Universitas Hasanuddin, Indonesia
[2]Department of Statistics, Universitas Negeri Makassar, Indonesia

Article Info

Article history:

Received : 07-30-2022
Revised : 10-23-2022
Accepted : 11-13-2022

Keywords:

Sentiment Analysis; Nave Bayes;
Lexicon-Based;
TikTok;
Google Play Store.

ABSTRACT

On TikTok application, there are several types of content in the form of education, cooking recipes, comedy, various tips, beauty, business, etc. However, some non-educational contents sometimes appear on TikTok homepage even though minors can access the app. As a result, TikTok application can influence the behavior of minors to be disgraceful, therefore, an assessment of the application can be one of the objects for conducting sentiment analysis. The purpose of this study is to compare the results of sentiment analysis on TikTok application using Naive Bayes with Lexicon-Based and without Lexicon-Based features. We used the TikTok reviews on Google Play Store as our data. According to the analysis, without Lexicon-Based feature, we obtained the accuracy rate, precision rate, and recall rate of 83%, 78%, and 69%, respectively. Meanwhile, the accuracy, precision, and recall rates using the Lexicon-Based feature were 85%, 91%, and 93%, respectively. Therefore, we concluded that sentiment analysis using Naive Bayes with Lexicon-Based feature was better than without Lexicon-Based feature on TikTok reviews.

Accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: https://doi.org/10.30812/varian.v6i1.2205

Corresponding Author:

Siswanto
Department of Statistics, Hasanuddin University
Email: siswanto@unhas.ac.id

A. INTRODUCTION

The Minister of Communication and Information said that Indonesia is a country with the fourth largest number of internet users in the world. This encourages the acceleration of information dissemination in cyberspace. In addition to being an information centre, internet services have built a new stigma, namely communicating without boundaries. Everyone can be connected to each other with the help of internet services. Along with the development of the internet which is increasingly fast and easy to access anywhere and anytime, the internet has become a fast means of communication. Unlimited communication is also supported by the rapid development of technology. The latest technology is very developed, there are many media that can be a means of communication. One of the modern technologies that are very helpful in communication is a smartphone. Smartphone is a mobile phone that can work almost the same as a computer device but with a version that is easy to carry everywhere. Smartphones with its features make it a necessity for most people. Smartphones can help humans with all their activities, such as as a learning tool where someone is able to browse the internet to get information related to lessons (Wilantika, 2015), can also read books through smartphones which are commonly called e-books. But most people make smartphones as a trend, lifestyle, prestige that can express their activities, interests and opinions (Pertiwi, 2020). Smartphones can also be used as a medium of entertainment.

Social media is one of the entertainment media that can be accessed on smartphones. Many social media that can be used to find entertainment such as Facebook, Instagram, Twitter, TikTok, etc. One of the most popular social media is TikTok. In 2020,

TikTok became a popular social media in Indonesia. TikTok is an application originating from China which was launched by Zhang Yiminy in 2016 (Wahyudi and Sibaroni, 2022). In addition to being an entertainment medium, TikTok also encourages the creativity of its users, because on TikTok users can create videos with a duration of 15 seconds until 3 minutes by using many features such as adding music, filters, stickers, and so on. On TikTok application, there are many types of content, such as educational content, cooking recipes, comedy, various tips, beauty, business, etc, hence it is not surprising why TikTok has become a popular social media. However, some non-educational content sometimes appears on the TikTok homepage even though all ages can easily access this application. Hence, TikTok could have a bad influence on minors because they have not been able to properly sort out what is good and right.

From the description above, the assessment of the TikTok application has varied. Thus, an assessment of TikTok application can be one of the object in conducting sentiment analysis. Sentiment analysis is textual data processing that aims to obtain information on a text. Various kinds of algorithms can be used in analyzing sentiment analysis such as Support Vector Machine (Bahri et al., 2022), Decision Tree C5.0, Classification and Regression Tree (Yusran et al., 2022). The information obtained is in the form of classification results with positive and negative categories. (Kundi et al., 2014) conducted the sentiment classification using Lexicon-Based framework which detected and scored the slangs used in the tweets. (Goel et al., 2016) performed real time sentiment analysis of tweets using Naive Bayes. (Le et al., 2019) showed the use of the dictionary as the input source improved the effectiveness of Naive Bayes by reducing the size of the training corpus and, as a result, training time. (Mustofa and Prasetiyo, 2021) conducted a research about sentiment analysis using Lexicon-Based on Twitter regarding the new normal campaign and obtained the accuracy of 79.72%. From several studies that have been conducted, the Naive Bayes method is the most frequently used method and produces a good accuracy. The purpose of this study is to compare the results of sentiment analysis on TikTok application using Naive Bayes with Lexicon-Based and without Lexicon-Based features. The novelty of the study is the sentiment analysis using Naive Bayes with Lexicon-Based feature was performed on TikTok user reviews on Google Play Store unlike the previous studies collected the data on Twitter.

B. LITERATURE REVIEW

1. Sentiment Analysis

Sentiment analysis, also known as opinion mining, is the process of automatically understanding, extracting, and processing textual data to obtain sentiment information that includes an opinion sentence. Sentiment analysis is used to determine whether a person tends to have negative or positive ideas or opinions about a situation or an object. Finding market trends and consumer attitudes about a product is one use of sentiment analysis in the real world. Research and applications based on sentiment analysis are growing quickly due to the scope of their influence and advantages (Choudhary and Choudhary, 2018).

Opinion mining is a branch of data mining where opinion mining generally assumes subjective information and classifies it into positive opinions or negative opinions. In general, opinion mining can be divided into several models (Setiawan et al., 2014). Sentiment analysis is a component of opinion mining, a method for comprehending, preprocessing (which involves decreasing data), and automatically processing textual data in order to gather information. Natural Language Processing (NLP) and text analysis are used in sentiment analysis to find and extract subjective information from a text (Gaur and Sharma, 2017). News can be used as research material from opinion mining/sentiment analysis because news contains more words than one posting on Facebook and Twitter. A news story has only one main review material that is the focus of its discussion.

2. Naive Bayes Classifier

Naive Bayes classifier is a statistical-based classification method based on the Bayes theorem to classify data into predetermined classes. It is called 'naive' because the value of an attribute has no effect on the value of other attributes which is called conditional independence. Naive Bayes classifier shows high accuracy and speed when implemented to large data compared to the performance of decision trees and some neural network classification algorithms. Equation (1) shows the general Naive Bayes classifier calculation.

$$P(c|w) = \frac{p(C|W)p(c)}{p(w)} \tag{1}$$

where $P(c|w)$ is the posterior or probability of class c against the word w, $p(C|W)$ is the likelihood or probability of the word w against class c, $p(c)$ is the prior or probability of occurrence of class c, and $p(w)$ is the evidence or probability of the occurrence of the word w. There is a calculation of the likelihood, prior, and evidence values in Equation (1). The calculation of the likelihood

value in Equation (2) uses the multinomial model. This model calculates the number of occurrences of each word in a document.

$$P(c|w) = \frac{count(w,c)}{count(c)} \tag{2}$$

where count (w, c) is the number of occurrences of words w in the class c and $count(c)$ is the number of occurrences of all words in the class c. The problem that is often found in multinomial model calculations is the presence of a word that never appears will result in a zero-value calculation which is called the zero frequency problem (Kikuchi et al., 2015). How to deal with this problem, do laplace smoothing. Laplace smoothing solves this problem by adding a value of 1 to the word or to the numerator so that it is considered to have appeared once and adding a unique word to the denominator. To calculate the prior is shown in Equation (3).

$$P(C_j) = \frac{N_C}{N} \tag{3}$$

where N_C is the number of documents in the training data of class c and N is the number of documents in the training data. To calculate the value of evidence is shown in Equation (4).

$$P(w) = \frac{|w|}{|s|} \tag{4}$$

where $|w|$ is the number of word w and $|s|$ is the number all the words that appear in the entire document.

3. Lexicon-Based Feature

Lexicon-Based feature is a feature or words that have been weighted based on a dictionary or lexicon. Weighting is performed for each word that includes positive or negative sentiments. The purpose of using Lexicon-Based feature is to determine the sentiment orientation of a word. A technique for sentiment analysis called Lexicon-Based makes use of a lexical or language source as a dictionary. The basis of this approach is to compare the terms in the sentiment dictionaries and determine how frequently they appear in the text (Catelli et al., 2022).

The sentiment value of a word is a real number with an interval of 0 to 1. If the sentiment value of a word is close to 1, then the word has a more positive sentiment. Meanwhile, if it is close to 0, then the word has a more negative sentiment. The sentiment value will later be integrated into the posterior calculation for each positive class and negative class. The calculation is shown in Equation (5) (Mehto and Indras, 2016).

$$P(c|w) = \frac{P(W|C) + senti\_score \times p(c)}{p(w) + senti\_score} \tag{5}$$

C. RESEARCH METHOD

This study used reviews regarding TikTok application on Google Play Store. By using Google-Play-Scrapper library in Python, we obtained 10000 data, but after the cleaning process, we obtained 1499 data. Data were consisted of variable User Name, Rating, Time, and Review. Table 1 shows some of the data used.

Table 1. The Data Used

No	User Name	Rating	Time	Review
1	Friska Manopo	4	31/05/2022 03.02	Min kenapa akun tiktok aku yg brnama mnpo_ tidak bisa dibuka padahal tidak ada maslaah apapun dan saya tidak mengepost video aneh" tolong ya min soalnya udah 2 bulan
2	Ezra Faris	5	30/05/2022 13.22	Tolong saya pengen fyp tik tok me:faris.gktau tapi g fyp g masalah karena apk ini bagus banget
3	Haria H	1	30/05/2022 13.12	ngga bisa belanja di tiktok live..alasannya percobaan terlalu banyak..padahal akun baru..heran
⋮				
10000	Jihan Atiyyah	2	13/11/2018 14.09	Padahal penyimpanan di hp ku gk seberapa penuh tapi pas mau bikin video malah keluar tulisan penyimpanan penuh, aku juga udah cek dan hapus beberapa video dan gambar di hp, ketika mau bikin video muncul tulisan jaringan tidak stabil padahal udah aku cek jaringan aku. Jadi tolong diperbaiki lagi ∂Y^{TM}?

We performed Naive Bayes with Lexicon-Based feature using software Python and R. The steps are shown the flowchart in Figure 1.

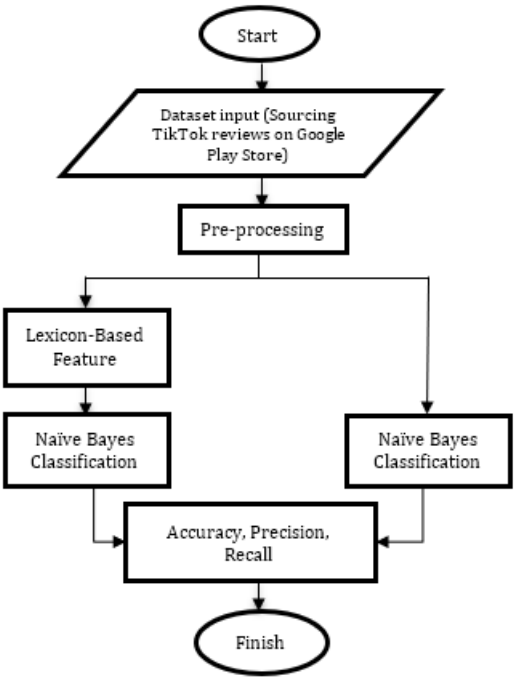

Figure 1. The Metodology Flowchart

The analysis to be carried out are as follows:

1. Naive Bayes Method with Lexicon-Based
   (a) Sourcing data on Google Play Store for TikTok application using text mining method.
   (b) Pre-processing data.
   (c) Performing Lexicon-Based analysis.
   (d) Analyzing the results.
2. Naive Bayes method without Lexicon-Based
   (a) Pre-processing data.
   (b) Performing sentiment analysis using Naive Bayes method.
   (c) Comparing the accuracy results of the Naive Bayes method with Lexicon-Based feature and without Lexicon-Based feature.
   (d) Drawing conclusion.

D. RESULTS AND DISCUSSION

1. Data Collecting

Collecting data was performed by scraping using Python on Google Collaboratory. The data obtained was 10,000 data, but the data was reduced to speed up the computing process therefore there were 1499 data left. The scraping data is shown in Table 2.

Table 2. TikTok App Review on Google Play Store

Rating	Review
1	sekarang mau upload video gabisa, harus uninstall dulu trus download lagi baru bisa upload
5	Bagus banget sama apk ini soalnya sudah diupdate gk nge bug
2	Kenapa efek kasat mata di TT saya tidak bisa digunakan

2. Data Preprocessing

a. Encoding labels

In this study, there are two categories of sentiment that are expected, namely positive and negative sentiments. Rating scores 1 and 2 will be categorized as negative analysis and ratings 4 and 5 as positive analysis. Additionally, data with a rating score of 3 will be eliminated because it is regarded as neutral or unknowable sentiment. The encoding results are presented in Table 3.

Table 3. Encoding Results

Rating	Review
Negative	sekarang mau upload video gabisa, harus uninstall dulu trus download lagi baru bisa upload
Positive	Bagus banget sama apk ini soalnya sudah diupdate gk nge bug
Negative	Kenapa efek kasat mata di TT saya tidak bisa digunakan

b. Cleansing

The text data obtained from the data collection does not fully contain the letters of the alphabet, it can be emoticons, symbols, or letters. Therefore, a cleansing process is needed. The cleansing results are presented in Table 4.

Table 4. Cleansing Results

Review	Cleansing
sekarang mau upload video gabisa, harus uninstall dulu trus download lagi baru bisa upload	sekarang mau upload video gabisa, harus uninstall dulu trus download lagi baru bisa upload
Bagus banget sama apk ini soalnya sudah diupdate gk nge bug	bagus banget sama apk ini soalnya sudah diupdate gk nge bug
Kenapa efek kasat mata di TT saya tidak bisa digunakan	kenapa efek kasat mata di TikTok saya tidak bisa digunakan

c. Filtering

Filtering was used to remove slang words, stop words and to do stemming. Most of the reviews used unstandardized Bahasa. In most cases such words do not affect significantly because they are not familiar to use. The Python programming language has a Sastrawi package which provides a feature to remove stop words and slang words from the data it has. The filtering results are presented in Table 5.

Table 5. Filtering Results

Review	Cleansing
sekarang mau upload video gabisa, harus uninstall dulu trus download lagi baru bisa upload	[sekarang, upload, video, harus, uninstall, dulu, download, lagi, baru, bisa, upload]
Bagus banget sama apk ini soalnya sudah diupdate gk nge bug	[bagus, sama, ini, sudah, update]
Kenapa efek kasat mata di TT saya tidak bisa digunakan	[kenapa, kasat, mata, saya, tidak, bisa, gunakan]

3. Sentiment Analysis

a. Visualization

Residual independence test is used to detect whether there is a correlation between lags. The following are the results of the residual independence test:

DOI: https://doi.org/10.30812/varian.v6i1.2205

Figure 2. Positive Sentiment

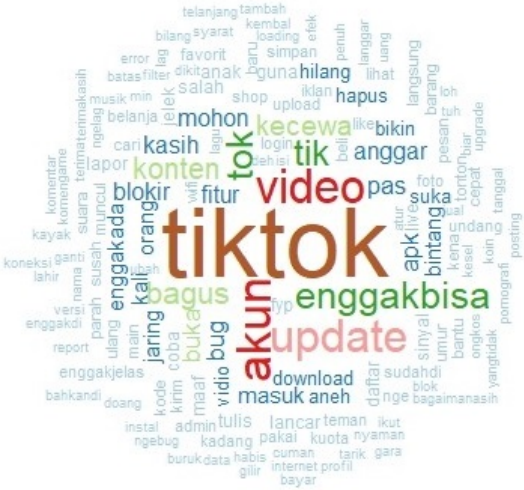

Figure 3. Negative Sentiment

Based on the word visualization, Figure 2 shows that the words with the most positive sentiments used in the reviews are bagus means good, video, TikTok, hibur means entertaint, suka means like. Meanwhile, Figure 3 shows that the words with the most negative sentiments used in reviews are akun means account, update, and enggakbisa means cannot.

b. **Naive Bayes without Lexicon-Based Feature**

With a dataset of 1499, the classification results obtained were negative sentiments of 34.69% and positive sentiments of 65.31%. Table 6 shows the accuracy of the Naive Bayes method without Lexicon-Based feature.

Table 6. The Test Results of Naive Bayes without Lexicon-Based Feature

Accuracy	Precision	Recall
0.83	0.78	0.69

Table 6 shows an accuracy of 83%. This shows that the percentage of testing data which class is classified correctly by the system is in accordance with the actual class. Then the compatibility value between the data classes generated by the

system and the actual class is 78% and the accuracy of the amount of data generated by the system is based on the actual class is 69%.

c. **Naive Bayes with Lexicon-Based Feature**

In the classification of TikTok review data, the cleaned dataset had 1499 data, then after performing the Lexicon-Based feature, the data was reduced to 1313 with the classification result of negative sentiment of 57.58% and positive sentiment of 42.42%. Table 7 shows the accuracy of the Naive Bayes method with Lexicon-Based feature.

Table 7. The Test Results of Naive Bayes with Lexicon-Based Feature

Accuracy	Precision	Recall
0.85	0.91	0.93

Table 7 shows an accuracy of 85%. This shows that the percentage of testing data which class is classified correctly by the system is in accordance with the actual class. Then the compatibility value between the data classes generated by the system and the actual class is 91% and the accuracy of the amount of data generated by the system based on the actual class is 93%.

E. CONCLUSION AND SUGGESTION

According to the analysis, without Lexicon-Based feature, we obtained the accuracy rate of 83%, precision rate of 78%, and recall rate of 69%. Meanwhile, with Lexicon-Based feature we obtained the accuracy rate of 85%, precision rate of 91%, and recall rate of 93%. Therefore, we concluded that sentiment analysis using Naive Bayes with Lexicon-Based feature was better than without Lexicon-Based feature on TikTok reviews on Google Play Store.

REFERENCES

Bahri, M. S., Hermawan, A., Kondy, E. P., Semida, R. J., et al. (2022). Performance comparison of supporting vector machine method without or with particle swarm optimization based on sentiment analysis whatsapp review. *International Journal of Academic and Applied Research (IJAAR)*, 6(6).

Catelli, R., Pelosi, S., and Esposito, M. (2022). Lexicon-based vs. bert-based sentiment analysis: A comparative study in italian. *Electronics*, 11(3):374.

Choudhary, M. and Choudhary, P. K. (2018). Sentiment analysis of text reviewing algorithm using data mining. In *2018 International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pages 532–538. IEEE.

Gaur, N. and Sharma, N. (2017). Sentiment analysis in natural language processing. *Int. J. Eng. Technol*, 3:144–148.

Goel, A., Gautam, J., and Kumar, S. (2016). Real time sentiment analysis of tweets using naive bayes. In *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, pages 257–261. IEEE.

Kikuchi, M., Yoshida, M., Okabe, M., and Umemura, K. (2015). Confidence interval of probability estimator of laplace smoothing. In *2015 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–6. IEEE.

Kundi, F. M., Khan, A., Ahmad, S., and Asghar, M. Z. (2014). Lexicon-based sentiment analysis in the social web. *Journal of Basic and Applied Scientific Research*, 4(6):238–48.

Le, C.-C., Prasad, P., Alsadoon, A., Pham, L., and Elchouemi, A. (2019). Text classification: Naïve bayes classifier with sentiment lexicon. *IAENG International journal of computer science*, 46(2):141–148.

Mehto, A. and Indras, K. (2016). Data mining through sentiment analysis: Lexicon based sentiment analysis model using aspect catalogue. In *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*, pages 1–7. IEEE.

Mustofa, R. and Prasetiyo, B. (2021). Sentiment analysis using lexicon-based method with naive bayes classifier algorithm on# newnormal hashtag in twitter. In *Journal of Physics: Conference Series*, volume 1918, page 042155. IOP Publishing.

Pertiwi, S. (2020). Pengaruh bauran pemasaran dan gaya hidup konsumen terhadap keputusan pembelian smartphone samsung (studi kasus pada program studi manajemen stie mikroskil medan). *Jurnal Wira Ekonomi Mikroskil*, 10(1):45–56.

Setiawan, K. Y., Hidayati, H., and Gozali, A. A. (2014). Twitter user opinion analysis at fine-grained sentiment analysis level toward public figure. In *E-Proceeding of Engineering*, volume 1, pages 639–646.

Wahyudi, D. and Sibaroni, Y. (2022). Deep learning for multi-aspect sentiment analysis of tiktok app using the rnn-lstm method. *Building of Informatics, Technology and Science (BITS)*, 4(1):169–177.

Wilantika, C. F. (2015). Pengaruh penggunaan smartphone terhadap kesehatan dan perilaku remaja. *Jurnal Obstretika Scienta*, 3(2).

Yusran, M., Rasyid, S., Sagita, E., Julia, R. N. D., et al. (2022). Sentiment analysis of sustainable development goals on twitter with classifying decision tree c5. 0 and classification and regression tree. *International Journal of Academic and Applied Research (IJAAR)*, 6(6).