Recursive Journal of Informatics
Vol. 2, No. 1, March 2024
p-ISSN 2963-5551 https://journal.unnes.ac.id/journals/index.php/rji e-ISSN 2986-6588

Comparison of Naive Bayes Classifier and K-Nearest Neighbor Algorithms with Information Gain and Adaptive Boosting for Sentiment Analysis of Spotify App Reviews

Meidika Bagus Saputro[1], Alamsyah[2]
[1,2] Computer Science Department, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang, Indonesia

Abstract. Currently, the development of technology is increase rapidly. One of the issues that appear with advance technology is data volume in the world has increase too. With the large data volumes that exist in the world it can be used to some purpose in many fields. Entertainment is one of the fields that have many interests from user in this world. Spotify is the example of entertainment apps that provided by Google Play Store to give online music streams to their users. Because that apps are provided by Google Play Store, many reviews of the user about the apps it can be classified to know the positive, negative, or neutral. One way to classified the review of user is make sentiment analysis. In this paper, to classify the review we use naïve Bayes classifier and k-nearest neighbors that will be compared with adding Information gain as feature selection and adaptive boosting as boosting algorithm of each classification algorithm that we used. The result of classification using naïve Bayes classifier with adding Information gain and adaptive boosting is 87.28% and k-nearest neighbor with adding information gain and adaptive boosting can perform accuracy of 80.35%.
Purpose: Knowing the result each of accuracy from the naïve Bayes classifier and k-nearest neighbor algorithm with adding information gain and adaptive boosting that we used and know how to doing the sentiment analysis step by step with the methods that chosen in this study.
Methods/Study design/approach: This study applied data preprocessing, lexicon based labelling with TextBlob, Normalization, Word Vectorization using TF-IDF, and classification with naïve Bayes classifier and k-nearest neighbor, information gain as feature selection, and adaptive boosting as boosting algorithm to boost the accuracy of classification result.
Result/Findings: The accuracy of naïve Bayes classifier with adding information gain and adaptive boosting is 87.28%. Meanwhile, by k-nearest neighbor with adding information gain and adaptive boosting reach the accuracy of 80.35%. This result obtained by using 60.000 dataset with data splitting 80% as data training and 20% as data testing.
Novelty/Originality/Value: Implementing information gain as feature selection and adaptive boosting as boosting algorithm to naïve Bayes classifier is prove that it can be increase the accuracy of classification, but not same when implementing in k-nearest neighbor. So, for the future research can applied another classification algorithm or feature selection to get better result.

Keywords: Sentiment analysis, Spotify, naïve Bayes classifier, k-NN, information gain, adaptive boosting.
Received May 10, 2024 / **Revised** July 07, 2024 / **Accepted** March 22, 2024

This work is licensed under a Creative Commons Attribution 4.0 International License.

INTRODUCTION

The development of technology and information in Indonesia from time to time continues which has a role and benefits for society in all aspects of life such as in the economic, educational, health, cultural, and other aspects of life. The development of technology and information making the high volume of data in the world increase too. However, increasing data that produced it can't to be comparable with the benefits and uses that can be used for various interests. Basically, the data collected from around that provide by the technology contains a lot of information that is can be used to determine opportunity and goals for some company, making decisions, and etc [1].

With all benefit from data that can help company, group, or individuals to achieve their goals, there is one method that can be used to properly process this data. Data mining is one useful new way in assisting an agency or company in finding important information and can be utilized for the benefit of the company [2]. Data mining is one of the best methods to processed data with big volume and fast. In data mining process,

[1]*Corresponding author.
Email addresses: meidikabagus@students.unnes.ac.id (Saputro)
DOI: 10.15294/rji.v2i1.68551

the processed data will be stored electronically and will be processed by a computer automatically using several techniques or certain method [3].

The development of technology and information prove that there are many activities at this time are supported by using digital technology. Technology really has an impact on several fields of work, one of them is the application in the entertainment. There are many types entertainment that can be accessed by the public available on a device technology such as music, video, images, etc. One of type entertainment which is much in demand by the society is music. Music is one of entertainment media that can express the feelings of its listeners through composition sound and singing produced [4]. One of the music platform that provide music in digital technology or streams for the user is Spotify [5]. Spotify is a music streams app that provides millions of songs, albums and originals podcast with a free service and a choice of paid services for para the user. Now, Spotify is available in Google Play Store platform with total 1 million downloads from around the world. With availability Spotify in Google Play Store, is possible to all user access and give some review about bad or good the apps when they using it. Many review that user give to the apps can be identify with one method that called sentiment analysis.

Sentiment analysis is a way to understand, process, and extract data in the form of text that aims to get sentiment information contained in a text or sentence [6]. Sentiment analysis can classify a sentence or represented by an opinion can be determine by positive or negative class [7]. So with sentiment analysis it can be help some people know about the apps is good or maybe bad to used. To doing classification for sentiment analysis can be classified by using classification algorithm.

In this study, naïve Bayes classifier and k-nearest neighbor is chosen for classify the sentiment analysis of Spotify app review. naïve Bayes classifier will be used in this study because the strength when classifying text has processing speed especially if you use a lot of data [8]. Meanwhile, k-nearest neighbor is chosen because it can classify the data from near distance of the data neighbor [9]. And then to make a better result, in this study using information gain for feature selection to choose top feature when classification processes and adaptive boosting as boosting algorithm for increase the result of the classification. Based on the description above, this study will focus on classification using naïve Bayes classifier and k-nearest neighbor algorithm with information gain and adaptive boosting for sentiment analysis of Spotify app review.

METHODS

This study focuses on comparing the accuracy of the naïve Bayes classifier and k-nearest neighbor classification algorithms for classification sentiment analysis. This accuracy comparison is accompanied by focusing on result of classification with classification algorithm, feature selection, and boosting algorithm. The process starts from input dataset, data preprocessing, labelling data, normalization, word vectorization with TF-IDF, feature selection, and classification using naïve Bayes classifier and k-nearest neighbor. The process stages can be seen in Figure 1.

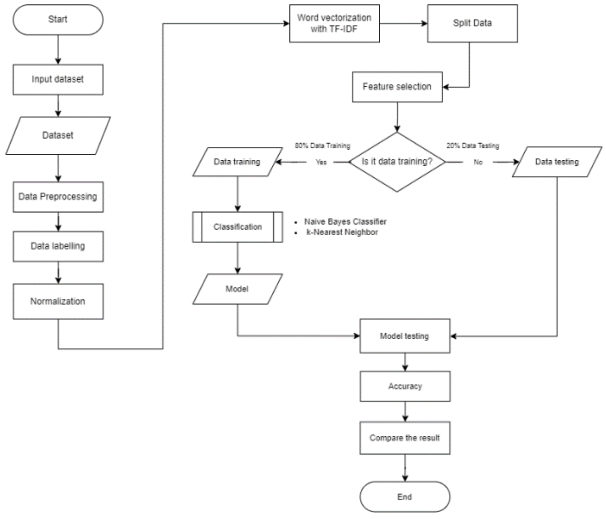

Figure 1. Flowchart of research processes

Dataset

In this study, dataset can be found in Kaggle website. This dataset is the result of web scrapping on Google Play Store about Spotify app reviews. For importance research this dataset can be downloaded via link https://www.kaggle.com/datasets/mfaaris/spotify-app-reviews-2022. This dataset has a total 60.000 data with three attributes. The sample of dataset can be seen in Table 1.

Table 1. Sample of dataset Spotify app review

time_submitted	review	rating
7/9/2022 15:00	Great music service, the audio is high quality and the app is easy to use. Also very quick and friendly support.	5
7/9/2022 14:21	Please ignore previous negative rating. This app is super great. I give it five stars+	5
7/9/2022 13:27	This pop-up "Get the best Spotify experience on Android 12" is too annoying. Please let's get rid of this.	4

Data preprocessing

Data preprocessing is stage that able to change data from unstructured to structured data [10]. This stage is the first stage of sentiment analysis classification. The purpose of this processes is managing the data, so it can be used for the next step. Data preprocessing consists of the following stages that can be explain below.

1. Case folding is first stage in data preprocessing that have purpose to equalizing or uniform all character by changing all the letters contained in a sentence to lowercase.
2. Cleansing is a second stage stage that is carried out with the aim of cleaning words or sentences that still contain noise. Words that contain noise and are removed in this process are words that have emoticons, hashtags, usernames, links, numbers, and several other characters [6].
3. Tokenization is next stage that carried out with the aim of cutting the strings in a sentence into a sequence of tokens that arrange the sentence [11].
4. Stopword removal is stage that have purpose to remove a word contained in a sentence [12]. This stage it can also be known as a collection of word lists that may not have a affect in carrying out a classification, for example words with affixes such as "with", "is", "that", etc [13].

Lexicon based labelling with TextBlob

Lexicon based is a method that can be used to classify a sentence that includes sentences with positive, negative, or neutral sentiment. Lexicon based labelling method is usually applied to research study that using machine learning algorithms in carrying out an analysis or classification [14]. One of implementation library using lexicon based is TextBlob. TextBlob is one of the libraries on lexicon based which used to provide labeling of a text automatically in a work related to text analysis. TextBlob performs automatic labelling based on the polarity or score of a data processed. TextBlob assesses the polarity of a data with a range of -1 to 1 [15]. The sentences that have polarity in range -1 till 0 is classified in negative class. Meanwhile, the sentences that has polarity in range 0 till 1 is classified in positive class.

Normalization

Normalization is a step that can be used to balance data to be processed to the next stage in order to get more balanced results. The purpose of the normalization stage is to produce balanced data based on comparative values between the before and after data and to form data with the same range. The normalization stage carried out in this study was to select data from the previous stage, in which case the labeling process obtained data output with three class categories, there are positive, neutral and negative. Based on the description before, to balance the data to be used in this study, there are data with positive and negative classes categories, it is necessary to carry out the normalization stage.

Split validation

Split validation is the next stage after conducting some processes by dividing the data into two, there are training data and testing data. In this study, the training data will be used to adjust the model in the analysis. And then, for testing data will be used as data to make predictions. In this study, comparation data that will be used is 80% for training data and 20% for testing data.

Word vectorization with TF-IDF

Word vectorization chosen in this study for determine the frequency value of a word in a document. In this process, the TF-IDF method is used to determine the relationship between words or terms and documents

by determining the weight of each word [16]. TF-IDF is chosen to estimate how representative a word in a document when compared to other words. At this stage the data which was originally in the form of text will be converted into a vector which then the vectorized data using TF-IDF will be used for the analysis or classification stage.

Feature selection with information gain

The feature selection is the stage to chosen and produce a more optimal result. In doing so, we need a technique that is used to sort the attributes from the highest to the lowest, namely by using the information gain technique. The feature selection process is carried out by preparing the TF-IDF data that has been obtained from the previous process, then the next thing to do is to calculate the total entropy value in the dataset, after that, the next step is to calculate the entropy of each feature in the dataset, and the last thing is reduce the total entropy value with the feature entropy. After the overall results are obtained, the results will be sorted from the largest to the smallest feature [17].

Classification

This stage is the stage for classifying the data that has been obtained in the previous stage. At this stage classification will also be carried out using the naïve Bayes classifier algorithm with information gain as feature selection and adaptive boosting as boosting algorithm and using the k-nearest neighbor algorithm with information gain as feature selection and adaptive boosting as boosting algorithm.

To classify the data in this study, each of method is have same step until get the accuracy. There are first step to classify is calculate the classification using classification algorithm. And then adding information gain as feature selection in each of classification algorithm. When adding information gain, can be define or determine k value to choose how many best features that will be used in classification with classification algorithm. Last, after got the accuracy of combination each classification algorithm and feature selection can be adding adaptive boosting as boosting algorithm for improve the accuracy from the previous accuracy obtained. To implement adaptive boosting, must be determine the estimators and learning rate to got the new accuracy for each combination methods.

The differs from the steps that have been carried out is when calculate using classification algorithm itself. When classify using naïve Bayes classifier the processes to get accuracy is by calculate the probability value [18]. Then, classify using k-nearest neighbor the processes is determine k value first and then accuracy will be obtained when calculate the closest neighbors appropriate with limitation of k value that determined before [9]. The stage of classification using naïve Bayes classifier with adding information gain and adaptive boosting can be seen in Figure 2.

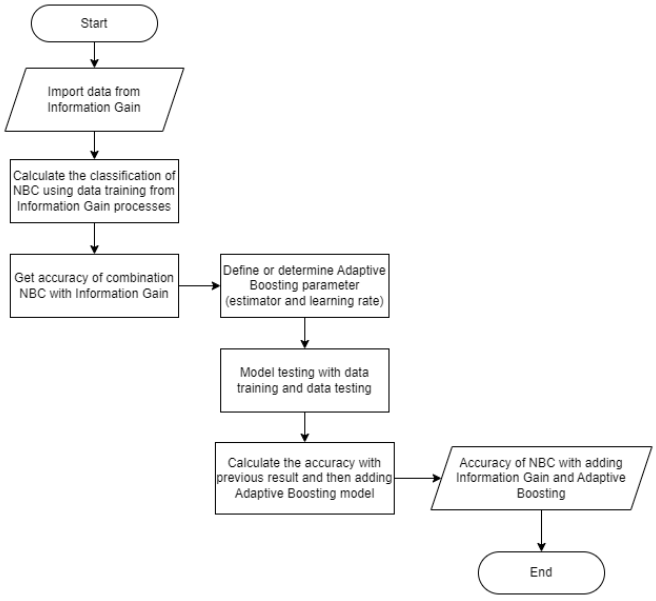

Figure 2. Classification of naïve Bayes classifier with adding IG and AB

And then, the stage of classification using k-nearest neighbor with adding information gain and adaptive boosting can be seen in Figure 3.

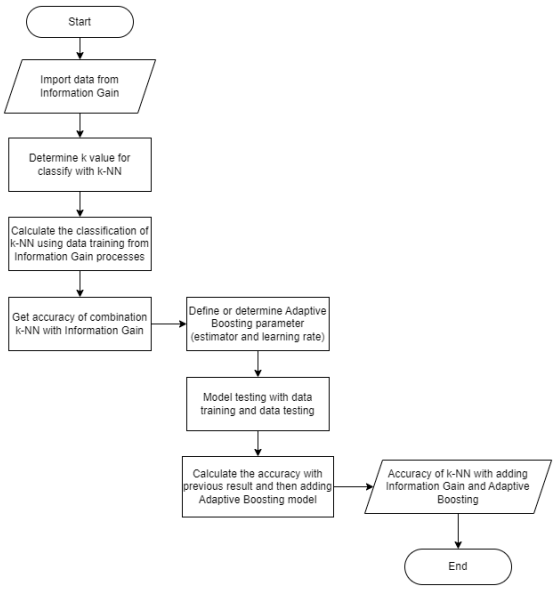

Figure 3. Classification of k-nearest neighbor with adding IG and AB

RESULT AND DISCUSSION

This study applies naïve Bayes classifier and k-nearest neighbor with adding information gain as feature selection and adaptive boosting as boosting algorithm. First processes to classify sentiment analysis of Spotify app review is data preprocessing. In data preprocessing, some steps is implemented in this study there are case folding, cleansing, tokenization, and stopword removal. The result of data preprocessing can be seen in Table 2.

Table 2. Result of data preprocessing

review	case folding	cleansing	tokenization and stopword removal	data after preprocessing
The music I've liked over my lifetime all in one place! Great App!!!	the music i've liked over my lifetime all in one place! great app!!!	the music ive liked over my lifetime all in one place great app	['music', 'ive', 'liked', 'lifetime', 'one', 'place', 'great', 'app']	music ive liked lifetime one place great app
Great device brilliant acsess to the music I like	great device brilliant acsess to the music i like	great device rilliant acsess to the music i like	['great', 'device', 'rilliant', 'acsess', 'music', 'like']	great device rilliant acsess music like
awesome app, love the podcasts!	awesome app, love the podcasts!	awesome app love the podcasts	['awesome', 'app', 'love', 'podcasts']	awesome app love podcasts

After get new data from preprocessing, the next process is to give label each of data that we have. In this study, labelling data using TextBlob that scoring or categorize sentiment each of data by polarity. The data that having polarity > 0 will give negative label, polarity = 0 will give neutral label, and then polarity > 0 will give positive label. The result of labelling with TextBlob can be seen in Table 3.

Table 3. Labelling data with TextBlob

review	polarity	sentiment
love definitely to recommend getting premium version but user friendly good way find new releases favorite artists	0.368560606	Positive
like best app music but dont like ads	0	Neutral
adds row premium really expensive	-1	Negative

Table 3 show every data that use in this study have each sentiment after labelling process using TextBlob. Next processes after labelling is normalizing data that selected two class of sentiment data, there are sentiment with class positive and negative. The objective of normalizing data is to balance data that will used in next processes. After that, the dataset will be divided for data training and data testing. The proportion of data training is 80% or 40.000 data and for data testing is 20% or 10.000 data.

From split validation that divide data for the classification, next processes is word vectorization using TF-IDF to calculate the term of document that we have in this study. In this processes data that is originally formed text will be change in vector for used in classification processes. The result of the accuracy using naïve Bayes classifier and k-nearest neighbor without any adding another method can be seen in Table 4.

Table 4. Accuracy classification with classification algorithm

algorithm	accuracy
naïve Bayes classifier	86.42%
k-nearest neighbor	84.24%

Then, to get better result can be implement information gain as feature selection in each of classification algorithm that will be used in this study. Usually for doing these processes, determine k value is must to choose the best feature on dataset that used for classification. In this study, k value for feature selection is 400 that means only 400 best feature of information gain result that will be used in classification processes. The accuracy of each classification algorithm with adding information gain as feature selection can be seen in Table 5.

Table 5. Accuracy classification with adding information gain

algorithm	accuracy
naïve Bayes classifier + information gain	84.44%
k-nearest neighbor + information gain	82.42%

From Table 5 know that the accuracy of each classification algorithm with adding feature selection in decreased. It can be happened because information gain only selects features that are considered important for classification and remove features that are considered unimportant. The removal of these features is considered to cause a decrease in accuracy [19]. So, to improve the accuracy of this classification, adaptive boosting as boosting algorithm is chosen for method addition to predict accuracy. The result of accuracy after adding adaptive boosting as boosting algorithm can be seen in Table 6.

Table 6. Accuracy classification with adding information gain and adaptive boosting

algorithm	learning rate	estimator	accuracy
naïve Bayes classifier + information gain + adaptive boosting	0.5	700	87.28%
k-nearest neighbor + information gain + adaptive boosting	0.7	800	80.35%

Based on the accuracy that obtained from each algorithm. Both combination algorithms produce novelty if compared with previous study. The combination of naïve Bayes classifier with information gain and adaptive boosting can be produce higher accuracy then previous study with same topic and object. Then, combination of k-nearest neighbor with information gain and adaptive boosting contribute novelty because there isn't any study with same topic and same object using this method that proposed in this study before. The comparison of this study with previous study can be seen in Table 7.

Table 7. Comparison with previous study

writer	algorithm	accuracy
Rahayu and Fauzi (2022)	naïve Bayes classifier	86.4%
	Support Vector Machine	84%
Daffa Rhajendra and Trianasari (2021)	naïve Bayes classifier	74.85%
Proposed method (2023)	naïve Bayes classifier + information gain + adaptive boosting	87.28%
	k-nearest neighbor + information gain + adaptive boosting	80.35%

The advantage of this study is showing that adding information gain and Adaptive boosting in classification algorithm to improve the accuracy of classification is proven by implementing the combination algorithm in naïve Bayes classifier. Improvement of this combination can be happened because the prediction of adaptive boosting in previous data classification is good then the accuracy can be increased.

Then, the disadvantage of this study can be seen in decreasing of accuracy when adding information gain in each of classification algorithm because the character of information gain that only selected the best feature and then remove the feature that considered not important when classifying. And then, decreasing accuracy when adding adaptive boosting in k-nearest neighbor and information gain is happen because the as know that characteristic of adaptive boosting that making the prediction by data from previous classification and this algorithm is sensitive with data noise.

CONCLUSION

Classification for sentiment analysis of Spotify app reviews using naïve Bayes classifier and k-nearest neighbor with the adding information gain as feature selection and adaptive boosting as boosting algorithms which are proven to produce good accuracy. With dataset that used in this study, combination of naïve Bayes classifier with information gain when k = 400 and adaptive boosting when estimator = 700 and learning rate = 0.5 can be produce best accuracy, there is 87.28%. Then, combination of k-nearest neighbor with information gain when k = 400 and adaptive boosting when estimator = 800 and learning rate = 0.7 can be produce the accuracy 80.35%. From the accuracy that obtain in this study, can be known that naïve Bayes classifier is get the better result than k-nearest neighbor. For the future research, can using same dataset with trying another combination of method such a classification algorithm, feature selection, and boosting algorithm.

REFERENCES

[1] D. S. Kusumo, M. A. Bijaksana, and D. Darmantoro, "Data mining dengan algoritma apriori pada Rdbms Oracle," *TEKTRIKA - J. Penelit. dan Pengemb. Telekomun. Kendali, Komputer, Elektr. dan Elektron.*, vol. 8, no. 1, pp. 1–5, 2016, doi: 10.25124/tektrika.v8i1.215.

[2] V. Moertini, "Data mining sebagai solusi bisnis, vol. 7, no. 1, pp. 44–56, 2017, [Online]. Available: https://eric.ed.gov/?id=ED539082%0Ahttp://www.win.tue.nl/~mpechen/research/edu.html.

[3] A. E. Pramadhani and T. Setiadi, "Penerapan data mining untuk klasifikasi penyakit ISPA dengan algoritma desicion tree," *J. Sarj. Tek. Inform. e-ISSN 2338-5197*, vol. 2, no. 1, pp. 831–839, 2014.

[4] D. Pangastuti, "Pengaruh musik dangdut terhadap perkembangan bahasa anak di TK Dharma Wanita Madiun 2014 / 2015," no. November, pp. 222–224, 2015.

[5] J. F. Andry and C. Tjee, "Analisis minat mahasiswa mendengarkan aplikasi musik berbayar dan unduhan musik gratis, Analysis of ttudent interest in listening to paid music applications and free music downloads," vol. 2, no. 2, pp. 9–15, 2019.

[6] G. A. Buntoro, "Analisis sentimen calon gubernur DKI Jakarta 2017 di Twitter," *Integer J.*, vol. 2, no. 1, pp. 32–41, 2017, [Online]. Available: https://t.co/jrvaMsgBdH.

[7] B. Liu, "Sentiment analysis: A multifaceted problem," *IEEE Intell. Syst.*, vol. 25, no. 3, pp. 76–80, 2010, doi: 10.1109/MIS.2010.75.

[8] R. Puspita and A. Widodo, "Perbandingan metode KNN, decision tree, dan naïve Bayes terhadap analisis sentimen pengguna layanan BPJS," *J. Inform. Univ. Pamulang*, vol. 5, no. 4, p. 646, 2021, doi: 10.32493/informatika.v5i4.7622.

[9] S. Surohman, S. Aji, R. Rousyati, and F. F. Wati, "Analisa sentimen terhadap review Fintech dengan metode naive bayes classifier dan k-nearest neighbor," *EVOLUSI J. Sains dan Manaj.*, vol. 8, no. 1, pp. 93–105, 2020, doi: 10.31294/evolusi.v8i1.7535.

[10] F. S. Jumeilah, "Penerapan support vector machine (SVM) untuk pengkategorian penelitian," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 1, no. 1, pp. 19–25, 2017, doi: 10.29207/resti.v1i1.11.

[11] E. Miana, A. Ernamia, A. Herliana, A. R. Sanjaya, and A. R. Sanjaya, "Analisis sentimen kuliah daring dengan algoritma naïve bayes dan k-nearest neighbor" vol. 4, no. 1, pp. 70–80, 2022.

[12] F. R. Irawan *et al.*, "Analisis sentimen terhadap pengguna Gojek menggunakan metode k-nearest neighbors, Sentiment analysis of Gojek users using k-nearest neighbor," vol. 5, no. 1, pp. 62–68, 2022, doi: 10.33387/jiko.

[13] A. Rachmat C and Y. Lukito, "Klasifikasi sentimen komentar politik dari Facebook Page menggunakan naive Bayes," *J. Inform. dan Sist. Inf. Univ. Ciputra*, vol. 02, no. 02, pp. 26–34, 2016.

[14] S. S. Salim and J. Mayary, "Analisis sentimen pengguna Twitter terhadap dompet elektronik dengan metode lexicon based dan k–nearest neighbor," *J. Ilm. Inform. Komput.*, vol. 25, no. 1, pp. 1–17, 2020, doi: 10.35760/ik.2020.v25i1.2411.
[15] S. Biswas, K. Young, and J. Griffith, "A comparison of automatic labelling approaches for sentiment analysis," pp. 312–319, 2022, doi: 10.5220/0011265900003269.
[16] M. Nurjannah and I. Fitri Astuti, "Penerapan algoritma term frequency-inverse document frequency (TF-IDF) untuk text mining mahasiswa S1 program studi Ilmu Komputer FMIPA Universitas Mulawarman dosen program studi Ilmu Komputer FMIPA Universitas Mulawarman," *J. Inform. Mulawarman*, vol. 8, no. 3, pp. 110–113, 2013.
[17] M. R. Maulana and M. A. Al Karomi, "Information gain untuk mengetahui pengaruh atribut," *J. Litbang Kota Pekalongan*, vol. 9, pp. 113–123, 2015.
[18] G. I. Webb, "Encyclopedia of machine learning and data science," *Encycl. Mach. Learn. Data Sci.*, no. April, 2020, doi: 10.1007/978-1-4899-7502-7.
[19] I. Kurniawati and H. F. Pardede, "Hybrid method of information gain and particle swarm optimization for selection of features of SVM-based sentiment analysis," *2018 Int. Conf. Inf. Technol. Syst. Innov. ICITSI 2018 - Proc.*, pp. 1–5, 2018, doi: 10.1109/ICITSI.2018.8695953.
[20] A. S. Rahayu and A. Fauzi, "Komparasi algoritma naïve Bayes dan support vector machine (SVM) pada analisis sentimen Spotify," vol. 4, pp. 349–354, 2022, doi: 10.30865/json.v4i2.5398.