SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 2, April 2023
DOI : https://doi.org/10.33395/sinkron.v8i2.12204

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Classification of Positive and Negative Sentiments Using the K-Nearest Neighbor Algorithm on iQIYI Application

Arief Pratama[1], **Susy Rosyida**[2]
[1,2]Universitas Nusa Mandiri, Jakarta, Indonesia
[1]ariefpratama3646@gmail.com, [2]susyrosyida@gmail.com

Submitted : Feb 14, 2023 | **Accepted** : Mar 3, 2023 | **Published** : Apr 1, 2023

Abstract: In the current state of the Covid pandemic, the government has implemented restrictions on community activities or PPKM, which has an impact on the number of cinemas in the country temporarily closed to reduce the spread of the virus. The number of films that have been postponed for release due to this outbreak and also the decreasing use of VCDs / DVDs have made movie streaming applications begin to be favored by the public, one of which is the iQIYI movie streaming application. iQIYI is a movie streaming app launched in April 2010, so that users can know that the iQIYI application is considered good is to do a sentiment classification on the application. Therefore, this study aims to implement sentiment classification in review data using the K-Nearest Neighbor (K-NN) algorithm. K-NN itself is an algorithm that functions to classify data based on its learning data (train data sets). The data used is iQIYI user reviews as many as 400 review data, the first stage carried out is the data cleaning process or Pre-Processing, the next step is to design a K-NN algorithm model in RapidMiner Studio software to process sentiment classification. The test results using 400 review data using the K-NN algorithm obtained an Accuracy value of 99.50% then a Precision value of 100% and a Recall value of 99.44%. Which means that this study managed to get the best and best algorithm in classifying positive reviews and negative reviews against the iQIYI application.

Keywords: iQIYI, K-Nearest Neighbor, Sentiment Prefix

INTRODUCTION

Since the covid-19 virus has become epidemic in various countries, including Indonesia, the government has immediately implemented a system of large-scale social restrictions, commonly called PSBB, which aims to reduce the spread of the covid-19 virus (Simarmata, Manik, Simanjorang, & Purba, 2022). The impact of the implementation of the PSBB is that many companies are forced to temporarily or permanently close their business activities, one of which is the closure of malls in Indonesia which also has an impact on film screening services (cinema) which ultimately makes film lovers feel disappointed (Cinephile) in Indonesia (Samudi, Widodo, & Brawijaya, 2020). Even though the spread of the Covid-19 virus has had many negative impacts, on the other hand, there are also positive sides, for example, the increasing use of the internet which goes hand in hand with the growing use of streaming movie applications along with the implementation of the PSBB which requires people to carry out activities at home (Al-shufi & Erfina, 2021).

The development of internet technology has contributed to expanding access and making the distribution of film streaming applications more and more at this time, the activity of watching movies is very easy to find in everyday life and can be enjoyed by anyone (Al Fath, Arini, & Hakiem, 2020). Of course, this has impacted the use of DVDs and VCDs which are starting to be abandoned gradually, including the DVD and VCD rental business which is decreasing and rarely found. The existence of films that are currently digitized makes films that used to have to provide a DVD player, now be easily operated. through a computer or laptop or even only through a smartphone. The many movie streaming applications in Indonesia such as iQIYI, Netflix, WeTV, Iflix, and so on provide various choices for film lovers (Cinephiles) in Indonesia (Wibowo, 2018).

An application certainly has its own advantages and disadvantages, where it can present various responses from application users such as satisfaction and disappointment with the application (Simarmata et al., 2022). The review column provided by Google Play store when a user installs one of these applications is a place to express

*name of corresponding author

user satisfaction and disappointment or opinions about the application (Indrayuni & Nurhadi, 2020). This can be used as material for classifying sentiment towards the iQIYI application, sentiment classification is used to find valuable information needed from unstructured data, so it is hoped that this research can determine the sentiment of iQIYI application users (Giovani, Ardiansyah, Haryanti, Kurniawati, & Gata, 2020).

iQIYI allows its users to provide reviews of their applications, these reviews are not only useful for potential application users but also useful for iQIYI application developers (Huang, Lv, & Sui, 2021), from the user's point of view, many positive or negative reviews can influence the decision to download the application or not, this is because the variety of available applications can confuse users to determine the best application (Shi & Zhou, 2021). From the developer's point of view, by reading user reviews, developers can find out what features need to be improved or fixed (Indriati & Ridok, 2016).

However, some of the existing app stores (App Stores) do not yet have a review feature so that users can rate the applications they want to download (Ai, 2020). Even though this feature is quite important so that users can choose the best movie streaming application because users can read previous user reviews before downloading it. Because previous user reviews are a series of opinions, facts, or observations on the application, the reviews themselves can be positive or negative reviews that can be used as consideration (Shiddieqy et al., 2016). From the explanation above, the authors are interested in conducting research on sentiment classification in reviews written by some users on the Google Play store regarding the iQIYI application (Vlassis, 2021), which will later be divided into two opinions, namely positive and negative using the K-Nearest Neighbor algorithm method which aims to obtain accuracy. the highest and the best results.

METHOD

The research method is a process or way that is chosen specifically in solving the problems posed in research. In general, research methods can be grouped into several types. The several types of research methods are as follows: Qualitative methods, quantitative methods, survey methods, facto exposure methods, and finally descriptive methods (Purnia & Alawiyah, 2020).

The descriptive method is a research method aimed at creating a systematic, actual, and accurate description of existing sample or population data. Descriptive research also does not require hypotheses or variable manipulation, because the symptoms and events already exist and the researcher just needs to describe them (Tanjung & Nababan, 2016). In addition, the authors carry out the process of collecting data in places that are the object of research, in supporting this research the researchers also use research instruments, and data collection techniques and determine the population to obtain the required number of samples, along with an explanation of each method of data collection carried out in writing this thesis.

RESULT

There are 922,000 review data on the play store for the iQIYI application. For the research sample, the researcher took as many as 400 review data after using the slovin formula to determine the research sample. The data taken is still in an unstructured form. The following are several steps to convert the data into structured data and are ready to be used in this study, namely:

Data Collection

The data collection process in this study uses the scraping method, the data will be mined from the Google Play Store website which provides the iQIYI application using webharvy software. Reviews of users of the iQIYI application mined as much as 400 data according to the number of research samples via the link https://play.google.com/tore/apps/details?id=com.iqiyi.i18n&hl=in&gl=US. The data that is mined is a list that is still unstructured, the data to be retrieved is the name, date, rating and comment.

Remove Duplicates

In the next stage, after we get 400 raw data, we remove duplicates. This stage aims to remove duplicate reviews so that there are no duplicate reviews. This process uses RapidMiner Studio version 9.10 software. In this process, we use operators in RapidMiner including: Operators Read Excel is used to read the specified excel file, Operators Remove Duplicates is used to remove duplicates from excel files by comparing all examples of each other based on specified attribute, The Examples Filter operator is used to select which examples from the excel file are saved and which examples are deleted.

Data Labeling

The result of removing duplicates obtained, and out of 400 review data, none of the data is the same and ready to be labeled. The purpose of this data labeling is not only to change the data to make it more structured but also to divide the dataset into two parts, namely training data and testing data. Training data is data that is

*name of corresponding author

used to train the system so that it can recognize the patterns you are looking for, while data testing is data that is used to test the results of the training that have been done before. The following is a dataset table that has been labeled.

Table 1. Data Labeling

Name	Date	Rating	Description
Dian Prasetya	20 July 2022	Rated 1 star out of 5 stars	Regret I upgraded iQIYI!!! So it's hard to watch the full episode and there are so many ads. It's better not to upgrade, it's still as good as the old one. If you make apps, you shouldn't be disappointed in use, your new program is very disappointing and to be honest, it's still good, the old program before I upgraded, I just regret it!!!
Bahrul Sopan	2 July 2022	Rated 5 stars out of 5 stars	Just installed, it and watched a movie for 1 hour without any ads. I read the reviews, how come there are many who complain, yes, some have even subscribed to complain about advertisements.

Table 1. above is an example of some raw data that has gone through the process of removing duplicates, the data is still unstructured. In order to make it structured data, it must first be labeled, the data will be labeled into a sentiment class to be used, there are two types of sentiment classes to be used, namely positive and negative

Table 2 Labeling Dataset

Review	Sentiment
Just installed it and watched a movie for 1 hour without any ads. I read the reviews, how come there are many who complain, yes, some have even subscribed to complain about advertisements.	POSITIF
It's good, but for free users like me, it's a bit less enjoyable to watch on iQIYI because there are too many ads, so my advice is to reduce the number of ads for non-VIP viewers.	POSITIF

DISCUSSIONS

After going through the process of removing duplicates and labeling datasets, the stages of data collection and data labeling have been completed. The next stage is the data pre-processing stage.

Pre-Processing

This pre-processing stage is the initial stage of text mining which aims to prepare text into data for later processing at a later stage, besides that the pre-processing stage is the stage where data is prepared to become data ready for analysis. There are several stages in this pre-processing, namely Case Folding, Cleansing, Tokenizing, Stopword, and Stemming.

Case Folding

The case-forming process is the stage where all uppercase characters are changed to lowercase letters. The stages of the implementation process using RapidMiner. The stages of the implementation process using RapidMiner and the results of the implementation are table 3 below.

Table 3. Result Case Folding

Before	After
please don't give ads until 5 characters are even worse per 30-minute ad, give it 1/2 so it's comfortable to watch it so that more users can make it easier, sorry for the free one, it's a good VIP clock, give it a little convenience for free, Is giving advertising comparable to the price of watching the film….I'll give 3 stars first if the advertising problem is fixed, I'll give 5 stars..that's all, thank you.......?	please don't give ads until 5 characters are even worse per ad 30 minutes give 12 so that it's comfortable to watch it so that more users can make it easier a little it's a pity that it's free, it's good that VIPs give it a little convenience that it's free to give ads it's not comparable to the price watch the film, I'll give 3 stars first if the advertising problem has been fixed, I'll give it a star. That is all and thank you
very good, I really like this application, but if I may, please reduce the ads reduce the ads, thank you.... iQIYI	very good, I really like this application, but if I may, I can reduce the ads, thank you, iQIYI

*name of corresponding author

Cleansing

The cleansing process is the process of reducing noise in the dataset, for example, characters that are removed such as tags (#), punctuation such as periods (.), commas (,), and other punctuation marks. The following is the implementation process with RapidMiner and the results are in table 4 below.

Table 4. Cleansing Results

Before	After
Too much buffering. Even though my signal and quota are okay. Already a VIP, there are still problems when watching it. Ad problem okay no problem. But please why is it buffering so much. It's been a long time, too.	too much buffering. even though my signal and quota are okay. I'm already VIP, but there are still problems when I watch it. ad problem okay no problem. but please why is it so often buffered. long time ago too.
Please, I'm already PREMIUM, but why is the loading so slow? It says feedback all the time, I'm tired of looking at it, the ad is still there when I pause the video, at this rate it's no different from the free acc, please fix it again, iQIYI.	please help, I'm already premium, but why is the loading so slow? it says feedback all the time, I'm tired of looking at it, the ad is still there when I pause the video, at this rate it's no different from the free acc, please fix it again, iQIYI.

Tokenizing

This tokenizing stage serves to separate or break one word from another, so each word will become a separate attribute. The following are the results of the tokenizing stage in table 5.

Table 5. Tokenizing Results

Before	After
Ah, how come I can't download iQIYI again, how come even though it's already a subscription place to watch but how come I'm being told to even feel forced to join VIP?	Ah, how come I can't download iQIYI anymore, how come even though it's already a subscription place to watch but how come I'm told to even feel forced to join VIP?
there are a lot of ads at a time wait 5 ads then most of the ads are 1550 seconds then can't be skipped actually the quality is good but yes the ads are	There are a lot of ads at a time, wait 5 ads, but most of the ads are 1550 seconds and can't be skipped, actually the quality is good, but the ads are

Stopword

In this fourth stage, the operator stopwords will remove unnecessary words such as the words "which", "and", "di", "is" and others. The following are the results of the stopwords stage in table 6.

Table 6. Stopwords Results

Before	After
This is why I want to renew my subscription, how come I can't usually use credit, but I keep getting rejected even though I have enough credit	renew subscriptions using credit are rejected even though the credit is sufficient
Hi, I'm having trouble playing the video, why is the screen only black and there's only sound, please fix this bug	having problems playing videos, black screen, sound, please fix it

Stemming

The last stage is stemming, which is a process to find the basic words of a word. By removing all affixes or affixes consisting of prefixes or prefixes, then suffixes or endings and also confixes or combinations of prefixes and suffixes in derived words. Stemming is used to change the form of a word into the basic word of that word which is in accordance with the proper and correct morphological structure of Indonesian. The following are the results of the stemming stage in table 7.

Table 7. Stemming Results

Before	After
adverts for seconds sometimes skip the name of the user for free, the ad for up to a minute disturbs the restlessness of the free user	Second ads sometimes disk up, free user names, ads for minutes, disturbing the comfort of free users
the application is like this, the advertisement is really cooking, the advertisement plays videos every minute, the advertisement for the application for viewing films, advertisements because users eliminate customer	an application like this, really cooks ads, plays video ads every minute, advertisements for movie viewing applications, advertisements because users disappear, customer satisfaction,

*name of corresponding author

satisfaction, advertisements are bad, avoiding advertisements just because it destroys user satisfaction, please consider other than that it's good	bad ads, avoid ads just because they destroy satisfied users, please consider other than good

TF-IDF weighting

The tf-idf weighting stage has the goal of giving a value or weight to the review sample which will be used as data testing in the calculation of the K-Nearest Neighbor algorithm. In this weighting, 10 data were taken from the dataset as a review and had 6 keywords, namely "advertising", "movie", "watching", "drama", "good" and "anime". After calculating the TF-IDF, the next step is to determine the first to tenth ranking by determining the weighted distance.

Implementation of the K-Nearest Neighbor Algorithm with Rapid Miner

This process uses RapidMiner tools and the K-Nearest neighbor algorithm which is in accordance with the initial goal. Testing this dataset to classify classes or positive and negative sentiments uses the K-NN algorithm method which is implemented with RapidMiner version 9.10 tools.

After the design stage of the K-NN algorithm model on RapidMiner, the next process is to measure the accuracy rate of the K-NN algorithm method which is evaluated using 10-fold cross validation with the number of k used in the K-NN algorithm being k = 5. The following are the results of testing the K-Nearest Neighbor classification algorithm method.

Test Results of the K-Nearest Neighbor Classification Algorithm Method

At this stage the researcher has obtained the results from testing the K-Nearest Neighbor algorithm method using 400 review data that have gone through the previous pre-processing stage, so that the data is clean from noise. The results of this test are displayed in the form of ROC Curves and Confusion Matrix.

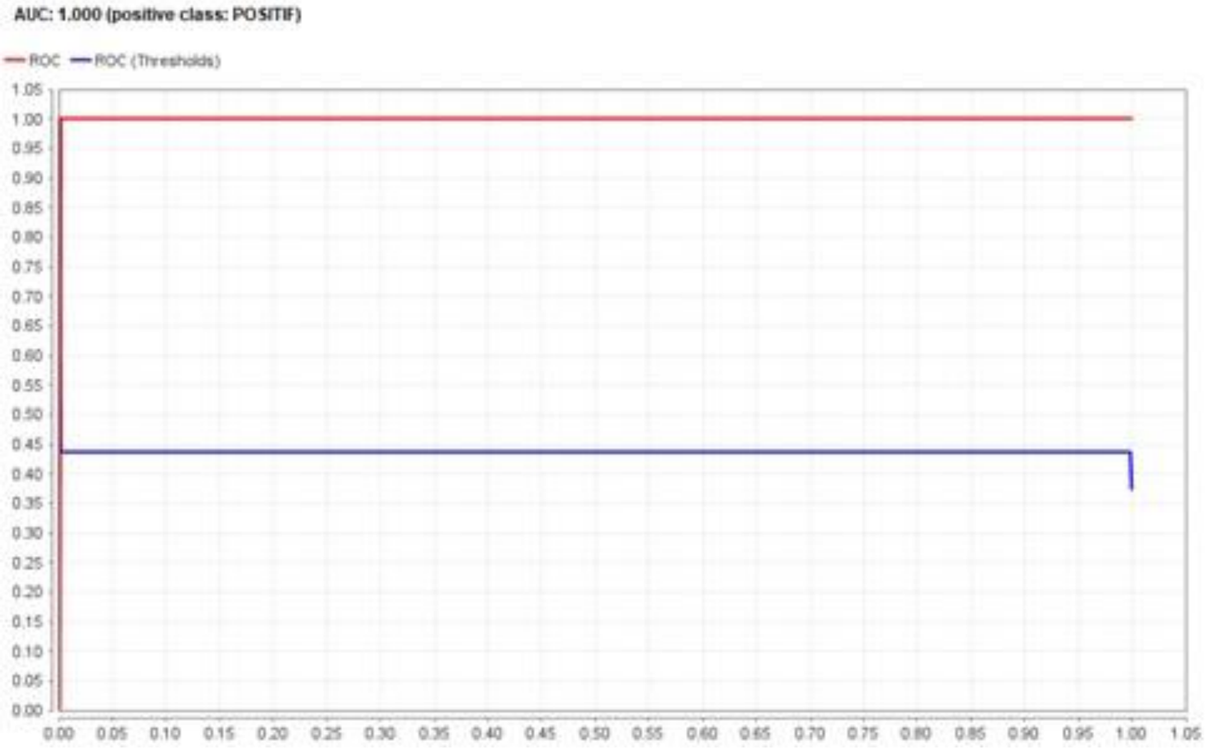

Figure 11 ROC curve

Table 8. Confusion Matrix

	True NEGATIF	True POSITIF	Class Precision
Pred. NEGATIF	45	2`	95.74%
Pred. POSITIF	0	353	100.00%
Class Recall	100.00%	99.44%	

The test results of 400 review data using the K-Nearest Neighbor algorithm get an accuracy value of 99.50% then a precision value of 100% and a recall value of 99.44%.

CONCLUSION

Based on the overall results of the research and discussion that has been carried out, it is concluded that research on the classification of positive & negative sentiments in the iQIYI application with the K-NN algorithm obtains an accuracy of 99.50%. With a Precision value of 100% and a Recall value of 99.44% using the RapidMiner Studio version 9.10 tools. The most sentiment results were positive sentiment, with 355 positive reviews and 45 negative reviews out of a total of 400 reviews. This indicates that more than 88% of iQIYI app users are happy with the app's performance.

REFERENCES

Ai, M. (2020). Research on the method of predicting the overvaluation of unicorn enterprises in China. *Academic Journal of Business & Management*, *2*(1), 14–24. https://doi.org/10.25236/AJBM.2020.020103

Al-shufi, M. F., & Erfina, A. (2021). Sentimen Analisis Mengenai Aplikasi Streaming Film Menggunakan Algoritma Support Vector Machine Di Play Store. *Sismatik*, 156–162.

Al Fath, M. K., Arini, A., & Hakiem, N. (2020). Sentiment Analysis Of Full Day School Policy Comment Using Naïve Bayes Classifier Algorithm. *SinkrOn*, *5*(1), 107–114. https://doi.org/10.33395/sinkron.v5i1.10564

Giovani, A. P., Ardiansyah, A., Haryanti, T., Kurniawati, L., & Gata, W. (2020). Analisis Sentimen Aplikasi Ruang Guru Di Twitter Menggunakan Algoritma Klasifikasi. *Jurnal Teknoinfo*, *14*(2), 115. https://doi.org/10.33365/jti.v14i2.679

Huang, Y., Lv, Z., & Sui, Z. (2021). Where Should Existing Video Streaming Platforms Improve: A Comparative Analysis of Netflix and IQiyi. *Proceedings of the 2021 International Conference on Public Relations and Social Sciences (ICPRSS 2021)*, *586*(Icprss), 585–592. https://doi.org/10.2991/assehr.k.211020.221

Indrayuni, E., & Nurhadi, A. (2020). Optimizing Genetic Algorithms for Sentiment Analysis of Apple Product Reviews Using SVM. *SinkrOn*, *4*(2), 172. https://doi.org/10.33395/sinkron.v4i2.10549

Indriati, I., & Ridok, A. (2016). Sentiment Analysis for Review Mobile Applications Using Neighbor Method Weighted K-Nearest Neighbor (Nwknn). *Journal of Enviromental Engineering and Sustainable Technology*, *3*(1), 23–32. https://doi.org/10.21776/ub.jeest.2016.003.01.4

Nurrun Muchammad Shiddieqy, H., Paulus Insap, S., & Wing Wahyu, W. (2016). Studi Literatur Tentang Perbandingan Metode Untuk Proses Analisis Sentimen Di Twitter. *Seminar Nasional Teknologi Informasi Dan Komunikasi*, *2016*(March), 57–64.

Purnia, D. S., & Alawiyah, T. (2020). *Metode penelitian strategi menyusun tugas akhir* (p. 58). p. 58.

Samudi, S., Widodo, S., & Brawijaya, H. (2020). The K-Medoids Clustering Method for Learning Applications during the COVID-19 Pandemic. *SinkrOn*, *5*(1), 116. https://doi.org/10.33395/sinkron.v5i1.10649

Shi, Y., & Zhou, J. (2021). *Analysis of Foreign Video Streaming Service Entering Chinese Streaming Media Market : A Case Study of Netflix*. *586*(Icprss), 337–343.

Simarmata, A. M., Manik, R., Simanjorang, O. C. R., & Purba, D. F. (2022). Data Mining using clustering method to predict the spread of Covid 19 based on screening and tracing results. *Sinkron*, *7*(4), 2355–2360. https://doi.org/10.33395/sinkron.v7i4.11740

Tanjung, H. S., & Nababan, S. A. (2016). Pengaruh penggunaan metode pembelajaran bermain terhadap hasil belajar matematika siswa materi pokok pecahan di kelas III SD Negeri 200407 Hutapadang. *Jurnal Bina Gogik*, *3*(1), 35–42.

Vlassis, A. (2021). Global online platforms, COVID-19, and culture: The global pandemic, an accelerator towards which direction? *Media, Culture and Society*, *43*(5), 957–969. https://doi.org/10.1177/0163443721994537

Wibowo, T. O. (2018). Fenomena Website Streaming Film di Era Media Baru: Godaan, Perselisihan, dan Kritik. *Jurnal Kajian Komunikasi*, *6*(2), 191. https://doi.org/10.24198/jkk.v6i2.15623