Sentiment Analysis on User Reviews Mutual Fund Ajaib App with K-Nearest Neighbor (KNN) Method

Adellia Maulanie Fadilla, Tacbir Hendro Pudjiantoro, Fajri Rakhmat Umbara and Asep Id Hadiana
Informatics Department, Faculty of Science and Informatics
Universitas Jenderal Achmad Yani, Cimahi, Indonesia
Adelliamaulanief09@gmail.com, thp@if.unjani.ac.id, fajri.rakhmat@lecture.unjani.ac.id, asep.hadiana@lecture.unjani.ac.id

Abstract

Online investment is an investment activity when viewed from the age range, these investors are dominated by the millennial generation. There are many examples of mutual fund online investment applications that have been widely downloaded by the public, for example, Ajaib. Sentiment Analysis is a technique for extracting text data to obtain information about positive, or negative sentiment. This research aims to analyze sentiment on user reviews of online investment applications, namely Ajaib by utilizing review analysis on the Google Play Store. Ajaib is a mobile application-based online stock and mutual fund investment platform that was simultaneously released in 2019, where its users are millennials. The many differences in online investment applications make Ajaib one of the superior applications in investing. This study aims to determine whether or not there are differences in features in terms of applications and services, prices, and promotions on interest in using the Ajaib application. The K-Nearest Neighbor method can help overcome the classification of opinions that are positive, or negative. After testing with K of 12, the highest accuracy results were obtained when K = 11 with an accuracy value of 85.0% with a precision of 90.2%, recall 87.5%, f1 value 88.8% with a tendency to also obtain reviews of the Ajaib mutual fund application tend to be positive.

Keywords
Online Investment, Mutual Fund, Sentiment Analysis, K-Nearest Neighbor

1. Introduction

There are various types of financial assets that investors can choose to invest their money. Included in the various types of financial assets. Mutual fund investment is often seen as an attractive investment for beginners, due to the relatively low and stable transaction costs. This can help investors to minimize risk and have greater certainty about the returns they may earn. Sentiment analysis usually refers to feelings, emotions, attitudes or opinions. Since textual data is used, there is a need to analyze the concept of expressing sentiment and calculating insights to explore businesses.By using the K-Nearest Neighbor method, researchers want to further examine the comparison of factors that influence interest in using the Ajaib application. Then classification will be carried out using the KNN (K-Nearest Neighbor) algorithm to determine whether the reviews on the data obtained are positive or negative.

1.1 Objectives

The formulation of the problem in this study is to compare whether the Ajaib mutual fund is better than previous research in terms of its accuracy value using TF-IDF extraction features, labeling using Lexicon Based and KNN grid search methods.

2. Literature Review

The use of the internet is able to influence people through the internet; we can get information, in addition, we can also give positive and negative opinions for certain reviews. By providing a lot of data or information on the internet, we use it for processing so, it will have new knowledge. Based on that, the author makes research, such as opinion classification by analyzing sentiment through a text mining approach, in this research; a method that is able to classify opinions accurately is needed. The scope of this research is the review of travel agents data processing using the K-Nearest Neighbor (K-NN) algorithm which uses 100 positive reviews and 100 negative reviews with six words related to sentiment, namely: Fast, Good, Great, Bad, Cencel, and Wait. It has evidence that by using the K-Nearest Neighbor (K-NN) algorithm, it achieves the best accuracy results and based on the calculations stated in the application. The accuracy point of reviewing travel agents using K-Nearest Neighbor (K-NN) algorithm has reached 87.00% and the AUC point is 0.916, the AUC point belongs to the Excellent Classification group so it is stated that K-Nearest Neighbor (K-NN) algorithm has reached 87.00%.(Ernawati and Wati 2018)

Online investment is an activity of investing capital either directly or indirectly with the hope that at some time the owner of the capital will get a number of benefits that are carried out online. There are examples of online investment applications that have been downloaded by many people according to Google Play Store, namely bibit and Bareksa. So, the purpose of this research is to analyze sentiment on user reviews of online investment applications, namely bibit and bareksa. The number of reviews that will be used in this research is 998 consisting of 484 positive sentiments and 514 negative sentiments for the bareksa application while for the bibit application using 1063 data consisting of 541 positive sentiments and 522 negative sentiments. The data also went through preprocessing and modeling stages. This research uses the CRISP-DM (Cross Industry Standard Process for Data Mining) model and the algorithm used in this research is K-Nearest Neighbors. Based on the results obtained from the modeling stage using the k-nearest neighbors algorithm and a ratio of 60:40 for training data and testing data, the precision and recall accuracy values generated from each application are for Bibit 85.14%, 91.91%, and 76.44% while for bareksa are 81.70%, 87.15%, 75.73%.(Adhi Putra 2021)

Nowadays, there are various kinds of financial instruments that can be used to add value to financial assets, one of which is stocks. Stocks are in high demand by the public because they offer a significant level of profit with relatively affordable capital. One way to benefit from stocks is to benefit from rising stock prices. There are various factors that can affect changes in stock prices, one of which is stock market sentiment. We can get this stock market sentiment from the financial news circulating. In this paper, we will discuss the use of string matching algorithms for scraping financial news and processing the scraped data into a stock sentiment. The results of this sentiment analysis can be used in determining investor opinions on certain stocks or assets.(Dharmastuti and Dwiprakasa 2017)

Technology implementation has moved quite widely in human life, especially in Indonesia. We can carry out various activities with the help of technology, one of them is about payments, which can now be done by applying for technological advances. In Indonesia, many digital payments are commonly used, such as GOPAY, DANA, and Shopee Pay, which are the main focus of this study. Expressions of satisfaction with the use of each platform are widely contained through the Twitter dataset, which later becomes a medium for data collection. Based on the testing method implemented in this study, we found that the K-Nearest Neighbor algorithm can provide better accuracy values than the Naïve Bayes algorithm. Based on the implementation of the K-Nearest Neighbor method, we later discovered that in Indonesia, the public generally gives a better rating in terms of user satisfaction with features and ease of application, which GOPAY, Shopee Pay, and DANA own, respectively.(Maharani and Triayudi 2022)

In this study, an attempt was made to classify sentiment analysis of tweets on Twitter on various opinions about usury on bank interest in Indonesia. To summarize people's views on usury on bank interest, text mining techniques are used, and data mining uses the K-NN classification algorithm to predict the labels in the dataset. The results show a K-NN Accuracy of 70.59%. The result for Precision is 69.87%. While the result for Recall K-NN is 62.32%. So it can be seen that the K-NN classifier is good enough to be used with social media datasets as it provides more accurate and precise predictions. In the future, we should use a larger and more complex dataset with an increase in the number of labels and a greater range of sentiments of usury tweets on bank interest and can include non-standard Indonesian languages. Indonesian non-standard languages (Rasenda Lubis and Ridwan 2020)

3. Methods

This study focuses on researching sentiment reviews of Magic mutual funds with a system that displays data, as well as the accuracy of K-Nearest Neighbor classification, where this research uses several steps, The step below shows the complete flow of the proposed methodology and can be explained with 4 steps.

Step 1:DataCollection.Datacollectionis required to be collect and prepare data for further analysis.

Step2: Data Preprocessing. Data preprocessing is required to be clean and prepare data for further analysis. Missing customer information should be completed or removed from the data. Stages of sentiment analysis of mobile app reviews application reviews begins with processing training data and test data with text pre-processing.(Gunawan et al. 2017)The main task of data preprocessing is to remove and resolve data noise so that the calculation results are optimized (Azhar et al. 2022).

Step3: Lexicon-based Data Labeling. Data labeling is the process of assigning labels to data that is data, with the aim of forming a classifier model. The processing stage of this research consists of two processes, namely importing positive and negative lexicon word dictionaries and breaking down reviews into fragments of words aimed at weighting or determining positive, and negative lexicon-based scores (Agrani et al. 2020).

Step4:Feature Extraction TF-IDF. Data in the form of words will be converted into numbers by doing the weighting process, the weights are calculated using the TF-IDF / TF-IDF method to improve the performance of methods that are useful for increasing accuracy and reducing computation time (Trstenjak et al. 2014)

Step5:K-Nearest Neightbor (KNN) is the simplest classification model that uses the approach of k nearest neighbors and assigns a class based on the most votes (Danades et al. 2017).

Step6: Confusion Matrix is a measurement of accuracy or performance for classification problems, where the output can be two or more classes. Confusion Matrix is described by a table of 4 different combinations of predicted and actual values. There are four terms that represent the results of the classification process in the confusion matrix: True Positive, True Negative, False Positive, and False Negative (Ting 2017).

4. Data Collection

Data scraping is done using the python programming language. Through the python programming language, datasets such as id, username, at, score, content, can be collected to be used as data to be processed. The data source of this research is reviews that come from user accounts on reviews of the magic mutual fund application.

5. Results and Discussion

5.1 Data Collection

The data used in this research is taken from a collection of reviews that come from public data on the Google Play Store. This review data is obtained using a data scrapping process with the phyton programming language. The data taken was 1001 data starting from October 15, 2021 to April 10, 2022. Examples of data that have been scrapped are in the table below (Table 1).

Table 1. Data Collection

Username	*Score*	*at*	*Content*
Rosahoka	5	2022-04-10 11:12:46	This application is good, easy to understand at the time of application.
aconliuw	5	2022-04-10 07:05:33	Good app easy to understand

5.2 Data Preprocessing

1. Case Folding: converts capital letters in a document to lowercase.
2. Remove Unused Characters: the process to remove all characters that are not used in modeling.
3. Stopword Removal: the process of removing Stopwords words such as in, to, from.
4. Stemming: Changing each word to its base word form.This is the process of finding the root word by removing all affixes attached to the word (Irfani 2020).
5. Removing Duplicates: From the initial cleaning results, duplicates may be found. So that the model built is not biased, duplicate removal is done.

6. Replace Slang Words: Changing slang words into standardized words, e.g. "cave" becomes "me".
7. Weighting words using lexicon
8. Build classification model with K-Nearest Neighbor method
9. Evaluate the model to be analyzed into information.

5.3 Lexicon Based Feature

Lexicon Based Features is a type of features based on knowledge or knowledge that focuses on obtaining a lexicon based on opinions from the text and then identifying the polarity of the lexicon polarity. Lexicon is a collection of terms that are known. This feature provides weight by requiring the help of a lexicon/dictionary to classify documents into positive sentiment or negative sentiment sentiment or negative sentiment (Ruslim et al. 2019).

5.3 Feature Extraction TF-IDF

Add propose in TF-IDF weighting there are 3 steps that must be passed including getting a vector from the results of the TF-IDF calculation, getting a value vector from the IDF calculation results and finally multiplying the results of the TF and IDF calculations. The dataset that will be used in the TF-IDF process is data that has gone through the preprocessing stage in Table 2.

Tabel 2. 1 TF

Feature	TF				
	Text 1	Text 2	Text 3	Text 4	Text 5
Aplikasi	1	1	0	0	0
Bagus	1	1	0	0	0
Sangat	1	0	0	0	0
Mudah	1	1	1	0	0
Kita	1	0	0	0	0
Mulai	1	0	0	0	0
Investasi	1	1	0	0	0
Fitur	0	1	0	2	0
Paham	0	1	0	0	0
Tambah	0	0	0	1	0
Fitur	0	1	0	1	0
Total	0	0	0	1	0
Jumlah	0	0	0	1	0
Bid	0	0	0	1	0
Ask	0	0	0	1	0
Biar	0	0	0	1	0
Makin	0	0	0	1	0
Lengkap	0	0	0	1	0
Rating	0	0	0	0	1
Kurang	0	0	0	0	1
Jadi	0	0	0	0	1
Tabel	0	0	0	0	1
Running	0	0	0	0	1
Trade	0	0	0	0	1
Sering	0	0	0	0	1
Error	0	0	0	0	1
Mohon	0	0	0	0	1

Segera	0	0	0	0	1
Baik	0	0	0	0	1

The length of the IDF vector adjusts to the number of features, the value entered into the vector is the result of the calculation. For example, to calculate the IDF vector on the word "very" is as follows:

$$IDF_{fitur} = \log\left(\frac{N}{n_{fitur}}\right) + 1$$
$$IDF_{fitur} = \log\left(\frac{5}{1}\right) + 1$$
$$IDF_{fitur} = \log(5) + 1$$
$$IDF_{fitur} = 2{,}60945$$

Description :
N : the number of texts in the data to be counted
n*fitur* : the number of occurrences of the word in the dataset
The value of N is the number of texts in the data to be calculated. For example, the number of texts used in the table above is 5 texts and nfitur is the number of occurrences of words in the dataset. because the word "very" only appears once in the third text, nfitur is worth 1. So the IDF on the word very is 2.60945. The same thing will be done by every other word which can be seen in Table 3.

Tabel 3. 1 IDF

Feature	IDF
Aplikasi	2,60945
Bagus	2,60945
Sangat	2,60945
Mudah	2,60945
Kita	2,60945
Mulai	2,60945
Investasi	2,60945
Fitur	2,60945
Paham	2,60945
Tambah	2,60945
Fitur	2,60945
Total	2,60945
Jumlah	2,60945
Bid	2,60945
Ask	2,60945
Biar	2,60945
Makin	2,60945
Lengkap	2,60945
Rating	2,60945
Kurang	2,60945
Jadi	2,60945
Tabel	2,60945
Running	2,60945
Trade	2,60945

Sering	2,60945
Error	2,60945
Mohon	2,60945
Segera	2,60945

TF-IDF Stage: This stage is the stage that performs the multiplication of the TF result vector and the IDF result vector so that the results of the TF-IDF weighting are in Table 4.

Tabel 4. 3 TF IDF

Feature	TF-IDF				
	Text 1	Text 2	Text 3	Text 4	Text 5
Aplikasi	2,60945	2,60945	0	0	0
Bagus	2,60945	2,60945	0	0	0
Sangat	2,60945	0	0	0	0
Mudah	2,60945	2,60945	2,60945	0	0
Kita	2,60945	0	0	0	0
Mulai	2,60945	0	0	0	0
Investasi	2,60945	2,60945	0	0	0
Fitur	0	2,60945	0	5,2189	0
Paham	0	2,60945	0	0	0
Tambah	0	0	0	2,60945	0
Fitur	0	2,60945	0	2,60945	0
Total	0	0	0	2,60945	0
Jumlah	0	0	0	2,60945	0
Bid	0	0	0	2,60945	0
Ask	0	0	0	2,60945	0
Biar	0	0	0	2,60945	0
Makin	0	0	0	2,60945	0
Lengkap	0	0	0	2,60945	0
Rating	0	0	0	0	2,60945
Kurang	0	0	0	0	2,60945
Jadi	0	0	0	0	2,60945
Tabel	0	0	0	0	2,60945
Running	0	0	0	0	2,60945
Trade	0	0	0	0	2,60945
Sering	0	0	0	0	2,60945
Error	0	0	0	0	2,60945
Mohon	0	0	0	0	2,60945
Segera	0	0	0	0	2,60945
Aplikasi	2,60945	2,60945	0	0	0

5.4 K-Nearest Neighbour (KNN) Classification

After obtaining the TF-IDF weighting results, the next stage is to carry out the classification process using the KNN method. In this process, the calculation of the K-Nearest Neighbor method to classify objects based on the training data that is closest to the test data. In other words, the K-Nearest Neighbor calculation functions to determine the closest distance. Data that has gone through pre-processing will become training data for data to be tested by measuring proximity to existing data. The distance calculation that will be used in this research is Euclidian Distance.(Dicki Pajri Yuyun Umaidah 2021)

First, the classification process is carried out using the K-Nearest Neighbor method, after the data is classified into review categories, the next step is to find the highest probability value to classify test data in the most appropriate category, the test data is a document from the review. There are two stages in the classification of documents to be used. The first stage is training on documents whose categories are already known, while the second stage is the process of classifying documents whose categories are not yet known.(Rasenda et al. 2020)

5.4 Confusion Matrix

Confusion Matrix is currently is a widely used tool in evaluating performance is a measurement of accuracy or performance for classification problems (Putri Fitrianti et al. 2019) where the output can be two or more classes. Confusion Matrix is described by a table of 4 different combinations of predicted and actual values. There are four terms that represent the results of the classification process in the confusion matrix: True Positive, True Negative, False Positive, and False Negative. At this stage, method testing with Confusion Matrix is carried out. Testing in this case is described by the accuracy testing stages of the K-Nearest Neighbor method. This test is carried out to determine the accuracy value in the classification of Ajaib application reviews on the system that has been created. The system being tested has a number of classes, namely 2 sentiments, namely positive and negative, including a multi-class classification system. This accuracy test is carried out using Confusion Matrix where a matrix of predicted results will be compared with the original class of input data. This test was carried out using test data of 20% of the total data of 1001 data, namely getting an accuracy of 85% in Figure 1..

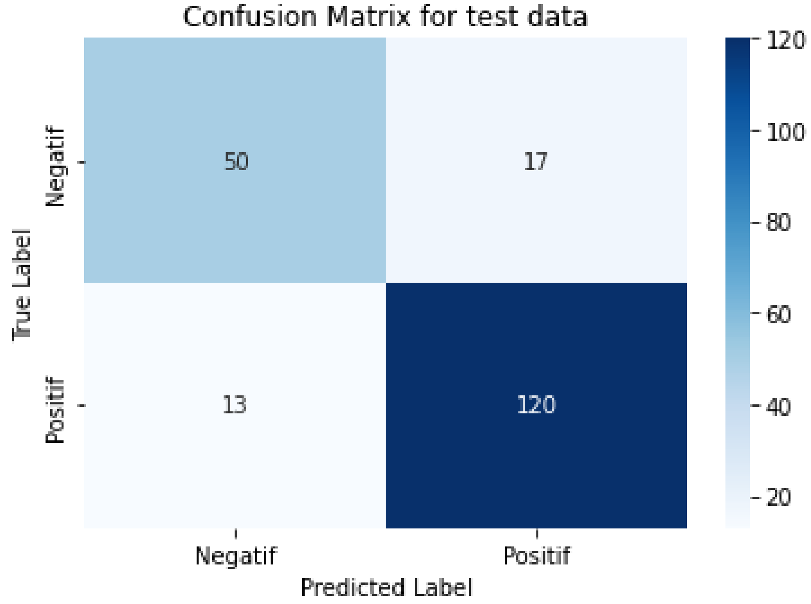

Figure 1. Confusion Matrix

6. Conclusion

Based on the results of the final project research, it can be concluded that the sentiment of users of the magic mutual fund application gives more positive reviews. The k-nearest neighbor classification modeler in this study gets an accuracy value (85%). The data used is 1001 datasets taken directly using the scrapping technique on Google Playstore. With a test accuracy level of Recall 87.6%, Presicion 90.2%, F-measurement 88.9% shows that the test accuracy results are greater using Presicion of 90.2%.

References

Amri, D., Pratama, D., Anggraini, D. and Anggriani, D., Comparison of Accuracy Level K-Nearest Neighbor Algorithm and Support Vector Machine Algorithm in Classification Water Quality Status, *Proceedings of the 2016 6th International Conference on System Engineering and Technology, ICSET 2016* (December 2016), pp. 137–41, 2017.

Anan, A. and Rikumahu, B., Perbandingan Analisis Sentimen Terhadap Digital Payment 'Go-Pay' Dan 'Ovo' Di Media Sosial Twitter Menggunakan Algoritma Naïve Bayes Dan Word Cloud Comparison of Sentiment Analysis Against Digital Payment 'Go-Pay' and 'Ovo' in Social Media Twitter Using N, *Agustus,* vol. 7, no. 2, pp. 2534, 2020.

Dharmastuti, Fara, C. and Dwiprakasa, B., Karakteristik Reksa Dana Dan Kinerja Reksa Dana Saham Di Indonesia, *Jurnal Ekonomi,* vol. 22, no. 1, pp. 94–116, 2017

Ferly, G., Fauzi, M. A. and Adikara, P. P., Analisis Sentimen Pada Ulasan Aplikasi Mobile Menggunakan Naive Bayes Dan Normalisasi Kata Berbasis Levenshtein Distance (Studi Kasus Aplikasi BCA Mobile), *Systemic: Information System and Informatics Journal,* vol. 3, no. 2, pp. 1–6, 2017.

Irfani, Fakhri, F., Analisis Sentimen Review Aplikasi Ruangguru Menggunakan Algoritma Support Vector Machine, *JBMI (Jurnal Bisnis, Manajemen, dan Informatika),* vol. 16, no. 3, pp. 258–66, 202.

Maharani, Putri, A. and Triayudi, A., Sentiment Analysis of Indonesian Digital Payment Customer Satisfaction Towards GOPAY, DANA, and ShopeePay Using Naïve Bayes and K-Nearest Neighbour Methods, *Jurnal Media Informatika Budidarma,* vol 6, no. 1, pp. 672, 2022.

Pajri, D., Umaidah, Y., Padilah, T. N., Implementation of K-Nearest Neighbor (K-NN) Algorithm For Public Sentiment Analysis of Online Learning, *Teknik Informatika dan Sistem Informasi* vol. 15, no. 2, pp. 121–30, 2021.

Putri Fitrianti, R., A-27 Implementasi Algoritma K-Nearest Neighbor Terhadap Analisis Sentimen Review Restoran Dengan Teks Bahasa Indonesia, *Seminar Nasional Aplikasi Teknologi Informasi (SNATi)*, pp. 1907–5022, 2019.

Putra, A., Dwiki, A.. Analisis Sentimen Pada Ulasan Pengguna Aplikasi Bibit Dan Bareksa Dengan Algoritma KNN, *JATISI (Jurnal Teknik Informatika dan Sistem Informasi),* vol. 8, no. 2, pp. 636–46, 2021.

Rasenda, Lubis, H. and Ridwan, R., Implementasi K-NN Dalam Analisa Sentimen Riba Pada Bunga Bank Berdasarkan Data Twitter, *Jurnal Media Informatika Budidarma,* vol. 4, no. 2, pp. 369, 2020.

Rizaldi, A., Surahman, A. and Juliane, C., Analisis Sentimen Terhadap Cryptocurrency Berbasis Python TextBlob Menggunakan Algoritma Naïve Bayes, *Jurnal Sains Komputer & Informatika (J-SAKTI),* vol. 6, pp. 267–81, 2022.

Ruslim, Ivana, K., Adikara, P. P. and Indriati, Analisis Sentimen Pada Ulasan Aplikasi Mobile Banking Menggunakan Metode Support Vector Machine Dan Lexicon Based Features, *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer,* vol. 3, no. 7, pp. 6694–6702, 2019.

Siti, E. and Wati, R., Penerapan Algoritma K-Nearest Neighbors Pada Analisis Sentimen Review Agen Travel, *Jurnal Khatulistiwa Informatika,* vol. 6, no. 1, pp. 64–69, 2018.

Ting, K. M., Confusion Matrix, *Encyclopedia of Machine Learning and Data Mining* (October), pp. 260–260, 2017.

Trstenjak, B., Mikac, S. and Donko, D., KNN with TF-IDF Based Framework for Text Categorization." *Procedia Engineering*, vol. 69, pp. 1356–64, 2014.

Biographies

Adellia Maulanie Fadilla is a final year undergraduate student in the department of informatics, JenderalAchmadYani University, Cimahi, Indonesia. Her primary interests are systems analysis, data and software engineering.

TacbirHendroPudjiantorois an Associate Professor. Doctoral Candidate from the Indonesian University of Education. Researcher in the field of Knowledge Management and handles several Information Systems projects.

FajriRakhmatUmbarais an Associate Lecturer. Received his Master's degree in Data Mining from Telkom University. He made books titled Cyber Ninja and Excel Intelligent and also seminar about shopping online and Microsoft Word for society.

Asep Id Hadiana received his Master's degree in Enterprise Information System from Indonesian Computer Univerity and a Doctor of Philosophy from Technical University of Malaysia Melaka. He is a lecturer in the

Informatics Department, Faculty of Science and Informatics, Universitas JenderalAchmadYani. Amongst his research interest are Cyber Security, Data Mining, Spatial Analysis, Location Based Services and Geographic Information Systems.