http://infor.seaninstitute.org/index.php/infokum/index
JURNAL INFOKUM, Volume 10, No.2, Juni 2022 ISSN : 2302-9706

SENTIMENT ANALYSIS COMPARE LINEAR REGRESSION AND DECISION TREE REGRESSION ALGORITHM TO DETERMINE FILM RATING ACCURACY

Rivaldo Sitanggang, Daniel Ryan Hamonangan Sitompul, Stiven Hamonangan Sinurat, Ruben, Andreas Stumorang, Denis Jusuf Ziegel, Julfikar Rahmad, *Evta Indra
Sistem Informasi, Fakultas Teknologi dan Ilmu Komputer, Universitas Prima Indonesia
Jl. Sampul No.3, Sei Putih Barat, Medan Petisah, Kota Medan

evtaindra@unprimdn.ac.id

Article Info
Received, 01 Juni 2022
Revised 20 Juni 2022
Accepted 22 Juni 2022

Abstract

Rating assessment in a film is the most important thing because it describes the satisfaction of film lovers with the films they have watched. With technological advances like now, we can easily find out the rating of a film by using a platform to accommodate the audience's review results, namely the Internet Movie Database (Imdb). The Machune Learning model that has been created can determine whether the film we watch is good based on ratings and reviews from moviegoers who share their experiences in watching similar films. Based on the results of the analysis of the two algorithms Linear Regression and Dicision Tree Regression, the best accuracy results from the Decision Tree Regression algorithm are 95.47%.

Keywords: Movie rating, IMDb, Linear Regression, Decision Tree Regression, Machine Learning.

1. INTRODUCTION

Rating rating in a film is the most important thing because it describes the satisfaction of film lovers with the films they have watched. With technological advances like now, we can easily find out the rating of a film by using a platform to accommodate the audience's review results, namely the Internet Movie Database (Imdb). Imdb is an online database related to information on people involved in making films, from actors/actresses, directors, writers to makeup artists and soundtracks.[1]. Many have used this platform as a reference in choosing which films are good or not.

There is a lot of information displayed on the Imdb site, one of which is film ratings, there are lots of film ratings ranging from the highest to the lowest. The film's rating is influenced by reviews from audiences from around the world. However, the drawback of this imdb is that it only shows trailers of films that will be shown in theaters to be judged by the reviews so that they make judgments based on non-objective opinions because those reviews do not watch the entire contents of the film, this will cause problems of inaccuracy in the assessment the rating given by the reviews to the films on the platform and the things that affect the reviews can judge a film just by watching the trailer[2][3]. The opinions given by the reviews need to be carried out by in-depth analysis by watching the contents of the entire film, to determine an objective assessment of a film by using sentiment analysis.

There have been many previous studies that have analyzed film ratings, for example “Movie Success Prediction using MachineLearning Algorithms and their Comparison” and “Extraction of Film Sentiment Reviews from Twitter with Naïve Bayes on Film Enthusiast Social Media Websites”[4][5]. This research uses data from imdb to analyze sentiment using a different algorithm. For the first study using the Random Forest Accuracy algorithm of 61%, AdaBoost Accuracy of 49.15%, Gradient Boost Accuracy of 56.68%, K-Nearest Neighbors Accuracy of 44.3% and Support Vector Machine Accuracy of 45.88%.

For the second study, the algorithm used Naïve Bayes Classifier Accuracy 79.95% and Rule-based System Accuracy 80.26%. The results of the first and second studies can be seen in the table below:

Table 1 comparison of the accuracy of the Sentiment Analysis Algorithm.

I . research algorithm	Accuracy	Research algorithm II	Accuracy
Random Forest	61%	Naive Bayes Classifier	79.95%
There is Boost	49.15%	Rule-based System	80.26%
Gradient Boost	56.68%		
K-Nearest Neighbors	44.3%		
support vector machine	45.88%		

In previous studies, the highest comparison accuracy value obtained was 80.26%, therefore it is necessary to develop a film review analysis to obtain greater accuracy than previous studies and state that the film rating is good or not based on the algorithm used. Because the greater the classification value and the accuracy of the sentiment analysis, the better[6].

Based on the background that has been described, the authors are interested in conducting a research entitled "Sentiment Analysis Comparing the Linear Regression Algorithm and Decision Tree Regression to Determine Film Rating Accuracy".

2. RESEARCH METHOD

2.1 Types of Research

In the world of entertainment, one of which is film requires an important role from film lovers, so that the film gets a good response or not by the general public. In this case, there is a need for a method to help film lovers or the general public who only know how to watch but cannot enjoy the films they watch. The method that will be made aims to help and make it easier for film lovers, especially ordinary people who don't know which films are good to watch, therefore researchers want to take advantage of Artificial Intelligence so that they can help and make it easier to determine the accuracy value and determine which films are good for the public to watch. .[7]. In this study, the algorithms used are Linear Regression and Decision Tree Regression from these two machine learning algorithms. The highest accuracy results will be sought to reduce classifier errors of a film rating and to determine the accuracy of the film so that it is precise.

*Linear Regression*is a data mining technique to determine that there is a relationship between the variable you want to predict with other variables[8].

*Decision Tree Regression*is the application of the most popular classification method today, this method is an item that can be grouped and modeled on a decision tree, so it can be easily understood[9].

2.2 Work Procedure

So that this research can run well and be completed on time, work procedures are made. The working procedures of this research are as follows:

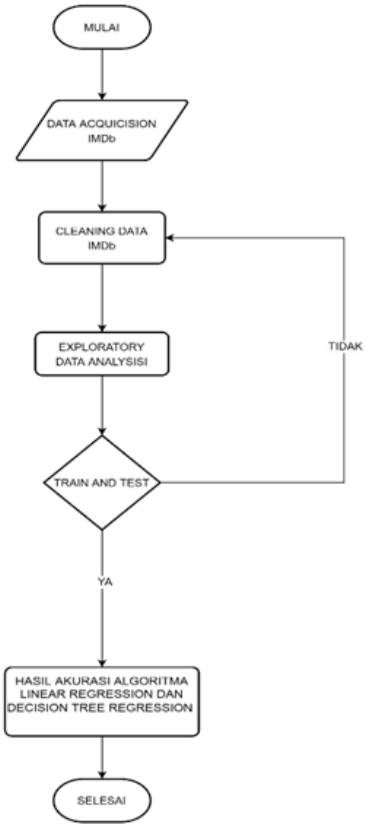


Figure 1 Flowchart

1. Data Acquisition

Data acquisition is a process carried out by researchers to collect data, in this case researchers obtain IMDb satay data from kaggle, the data contains 5043 rows and 28 columns.[10].

2. Data Cleaning

After the data is obtained, the next process is to delete data (cleaning data) this aims to remove empty data so that the data obtained is clean and makes the results more reliable and can be used by the model.[11].

3. Exploratory Data Analysis

*Exploratory Data Analysis*is a data exploration process that aims to understand the content and components of the data. In this study, EDA was carried out on the dataset to see the content of the data, the correlation of the data, the distribution of the data. Boxplot visualization is used to view the description of the dataset in a visual form, so that we can better understand how to divide the quartiles to outliers in the dataset.[12].

4. Train and Test

*Train and Test*is a process in which the dataset is processed using the Linear Regression and Decision Tree Regression algorithms to perform sentiment analysis of imdb film ratings using the python google colab programming language, after processing the data and obtaining accuracy results it will be seen whether the results obtained are good enough in sentiment analysis imdb. If the processed data does not get good results in conducting sentiment analysis, it will return to the data cleaning stage so that the data obtained is better than previous research[13][14].

5. Result Accuracy

The final result of machine learning data processing using the Linear Regression and Decision Tree Regression algorithms is an accuracy comparison for film rating analysis so as to get accurate results based on the algorithm used[15].

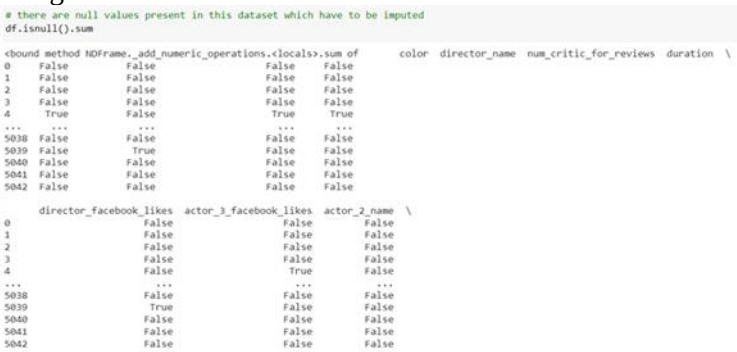
3. RESULTS AND DISCUSSION

3.1 Data Cleaning

At the data cleaning stage or data deletion which aims to clean the data that will be used by the model for processing so as to produce good output results. In this study, the data cleaning carried out is as follows:

3.1.1 Calculating Zero Value Data

At this stage the data with a value of zero is calculated so that the results of data processing carried out by the model can produce accurate accuracy. It can be seen in Figure 2 the process of calculating data that is zero and its results.



```
# there are null values present in this dataset which have to be imputed
df.isnull().sum
```

```
<bound method NDFrame._add_numeric_operations.<locals>.sum of      color  director_name  num_critic_for_reviews  duration  \
0    False          False                   False     False
1    False          False                   False     False
2    False          False                   False     False
3    False          False                   False     False
4     True          False                    True      True
...    ...            ...                     ...       ...
5038 False          False                   False     False
5039 False           True                   False     False
5040 False          False                   False     False
5041 False          False                   False     False
5042 False          False                   False     False

      director_facebook_likes  actor_3_facebook_likes  actor_2_name  \
0                       False                   False         False
1                       False                   False         False
2                       False                   False         False
3                       False                   False         False
4                       False                    True         False
...                       ...                     ...           ...
5038                    False                   False         False
5039                     True                   False         False
5040                    False                   False         False
5041                    False                   False         False
5042                    False                   False         False
```

Figure 2. The process of calculating data is zero

3.1.2 Calculating Film By Color

At this stage, films that have many colors and films that have two colors are calculated and the results are colored films with a value of 4815 films while films with black and white are 209 films which can be seen in Figure 3.

```
# We have 4815 colour movies and 209 Black and white movies.
df['color'].value_counts()
```

```
Color              4815
 Black and White     209
Name: color, dtype: int64
```

Figure 3. Calculation of film color

3.1.3 Calculating Movies by Language

At this stage, a film count was carried out based on the most languages used in the film, based on the dataset, the results of the calculation of films using English got the highest number of films, namely 4704 films which can be seen in Figure 4.

```
# A lot of movies in this dataset are in er
df['language'].value_counts()
```

```
English       4704
French          73
Spanish         40
Hindi           28
Mandarin        26
German          19
Japanese        18
Cantonese       11
Russian         11
Italian         11
Portuguese       8
Korean           8
Arabic           5
Danish           5
Hebrew           5
Swedish          5
Polish           4
Norwegian        4
Persian          4
Dutch            4
Thai             3
Chinese          3
Icelandic        2
None             2
Indonesian       2
Aboriginal       2
```

Figure 4. Film calculation by language

3.2. Results of Exploratory Data Analysis

At the Exploratory Data Analysis (EDA) stage in this study, it will be divided into 5 parts, namely visualizing the imdb score, visualizing the number of languages used, visualizing the name of the director (director), visualizing the country barplot and imdb score, and visualizing the dataset distribution barplot.

The first part of this visualization is showing the imdb score which is shown in Figure 5. The data set used in this study can be concluded that the highest rating from imdb has a score of 6.7 and has a percentage of 11.74% and the lowest score on imdb is 6.1 and a percentage of 9.43% .

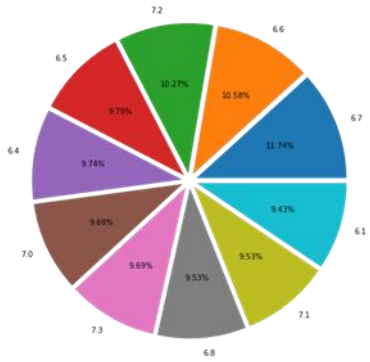


Figure 5. Imdb score diagram

In the second part, the visualization of the language used in imdb can be seen in Figure 6. Based on the data, the most widely used language is English for making films and the least used language is Japanese.

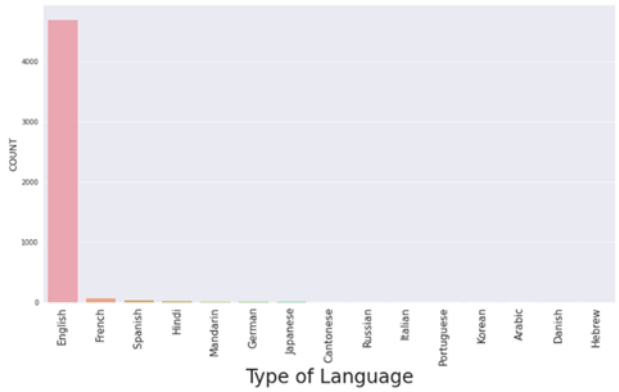


Figure 6. Visualization by language type

In the third part, the visualization of the name of the film director (film director) can be seen in Figure 7. Based on this data, it can be concluded that the director named Steven Spielberg has made more than 25 films.

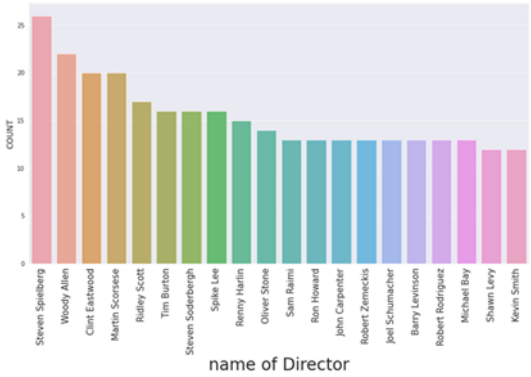


Figure 7. Name of Film Director

In the fourth part, the visualization of the country barplot and imdb scores are seen in Figure 8. Based on the barplot, it can be concluded that Kyrgyzstan has the highest IMDB score of any other country, in other words, the country has the most moviegoers from many countries.

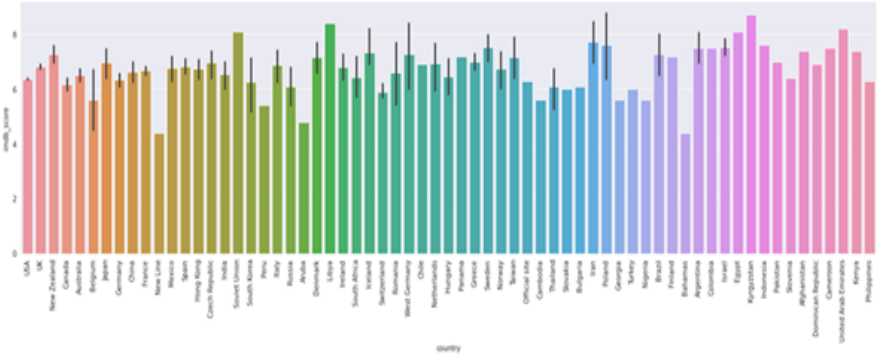


Figure 8. Barplot of Country and Imdb Score

In the fifth part, the visualization of the distribution of the entire dataset can be seen in Figure 9. Based on the barplot of the distribution of the satay data, it can be concluded that the data obtained are director_name, num_critic_reviews, duration, director_facebook_likes, actor_3_facebook _likes, gross(gross), genres, budget, country, title_year , imdb_score and movie_facebook_likes.

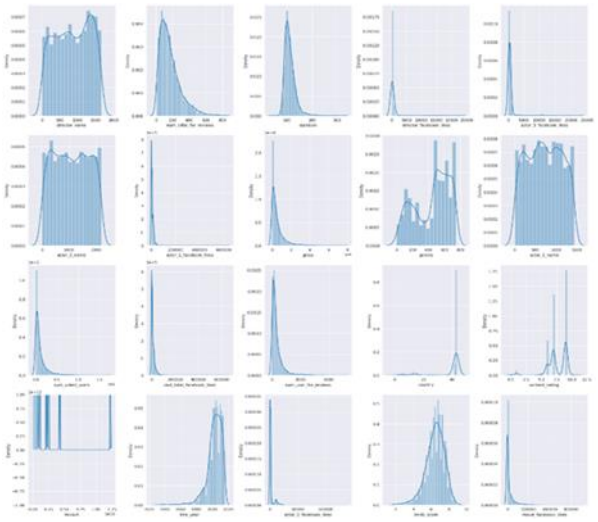


Figure 9. Barplot of dataset distribution

3.3 Splitting Dataset

Splitting dataset is the stage of separating variables into two parts, the first part is called dependent and the second part is independent, this is done to train and test. Splitting the dataset is done using the python library scikit-learn. From the results of the train and test, the results of the distribution of 80:20 are obtained with a random state of 40 which can be seen in Figure 10.

```
# Train_test_split
X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.2,random_state=40)
print(X_train.shape,X_test.shape,Y_train.shape,Y_test.shape)
# Splitting data set into training and testing.
```

```
(3004, 19) (752, 19) (3004,) (752,)
```

Figure 10. Train and Test Splitting dataset

3.4 Results of Modeling

3.4.1 Linear Regression

The results of making the model/training model from the Linear Regression algorithm have a Root Mean Square Error (RMSE) training data value of 0.12 or 12%, while the Root Mean Square Error (RMSE) test data is 0.11 or 11% and the accuracy value of the Linear Regression algorithm is 0.11 or 11%. 95.5%.

```
lm=LinearRegression()
lm = lm.fit(X_train,Y_train)
#Traindata Predictions
train_pred = lm.predict(X_train)
#testdata predictions
test_pred = lm.predict(X_test)
RMSE_test = np.sqrt(mean_squared_error(Y_test, test_pred))
RMSE_train= np.sqrt(mean_squared_error(Y_train,train_pred))
print("RMSE TrainingData = ",str(RMSE_train))
print("RMSE TestData = ",str(RMSE_test))
print('-'*50)
print('RSquared value on train:',lm.score(X_train, Y_train))
print('RSquared value on test:',lm.score(X_test, Y_test))
```

```
RMSE TrainingData =  0.12041595568987472
RMSE TestData =  0.11861105307571114
--------------------------------------------------
RSquared value on train: 0.40800950550852266
RSquared value on test: 0.4021001552231892
```

Figure 11. Root Mean Square Error (RMSE) training data and linear regression algorithm test data

```
[ ] errors = abs(test_pred - Y_test)
    # Calculating errors for using error values in mean absolute percentage error
```

```
[ ] # Calculate mean absolute percentage error (MAPE)
    mape = 100 * (errors / Y_test)
    # Calculate and display accuracy
    accuracy = 100 - np.mean(mape)
    print('Accuracy:', round(accuracy, 2), '%.')

    Accuracy: 95.5 %.
```

Figure 12. Linear Regression accuracy results

3.4.2 Decision Tree Regression

The results of making the model/training model from the Decision Tree Regression algorithm have a Root Mean Square Error (RMSE) training data value of 0.08 or 8% while the Root Mean Square Error (RMSE) test data is 0.12 or 12% and the accuracy value of the Decision Tree algorithm Regression of 95.47%.

```
DT=DecisionTreeRegressor(max_depth=9)
DT.fit(X_train,Y_train)
#predicting train
train_preds=DT.predict(X_train)
#predicting on test
test_preds=DT.predict(X_test)
RMSE_train=(np.sqrt(metrics.mean_squared_error(Y_train,train_preds)))
RMSE_test=(np.sqrt(metrics.mean_squared_error(Y_test,test_preds)))
print("RMSE TrainingData = ",str(RMSE_train))
print("RMSE TestData = ",str(RMSE_test))
print('-'*50)
print('RSquared value on train:',DT.score(X_train, Y_train))
print('RSquared value on test:',DT.score(X_test, Y_test))
```

```
RMSE TrainingData =  0.08302711763715252
RMSE TestData =  0.12485226930211782
--------------------------------------------------
RSquared value on train: 0.7185595074222233
RSquared value on test: 0.3375227054465846
```

Figure 13. Root Mean Square Error (RMSE) training data and test data of the Decision Tree Regression algorithm

```
[ ] errors = abs(test_preds - Y_test)
    # Calculating errors for using error values in mean absolute percentage error
```

```
[ ] # Calculate mean absolute percentage error (MAPE)
    mape = 100 * (errors / Y_test)
    # Calculate and display accuracy
    accuracy = 100 - np.mean(mape)
    print('Accuracy:', round(accuracy, 2), '%.')
```

```
Accuracy: 95.47 %.
```

Figure 14. Accuracy Decision Tree Regression Results

3.5 Algorithm Comparison

After completing the model creation and the results of model fitting are obtained, the model will be compared based on the Root Mean Square Error (RMSE) and the accuracy of the algorithm used. In this study, it can be concluded that the Linear Regression algorithm has a Root Mean Square Error Train (RMSE) with a value of 12%, a Root Mean Square Error Test (RMSE) with a value of 11% and an accuracy of Linear Regression of 95.5%, while Decision Tree Regression is based on Root The Mean Square Error Train (RMSE) has a value of 8%, the Root Mean Square Error Test (RMSE) has a value of 12% and the accuracy value of the Decision Tree Regression algorithm is 95.47%. The comparison visualization of the model can be seen in table 2.

Table 2. Comparison of the Linear Regression Algorithm and Decision Tree Regression Algorithm models.

Algorithm	RMSE Train	RMSE Test	Accuracy
Linear Regression	12%	11%	95.5%
Decision Tree	8%	12%	95.47%

4. CONCLUSION

With this research, it can help film lovers in determining which films are good to watch based on the Accuracy Rating of the film by using Machine Learning based on a classification algorithm, so

that with the film parameters on IMDb Plat From, a film review provider that is widely used by film lovers around the world. parts of the world. The Machune Learning model that has been created can determine whether the film we watch is good based on ratings and reviews from film connoisseurs who share their experiences in watching similar films. Based on the results of the analysis of the two algorithms Linear Regression and Dicision Tree Regression, the best accuracy results from the Decision Tree Regression algorithm are 95.47%.

REFERENCE

[1] N. A. Mayangky, D. N. Kholifah, I. Balla, and I. J. Thira, "Pengaruh Rating Film Terhadap Jumlah Audience Yang Menonton Film," *Indones. J. Softw. Eng.*, vol. 5, no. 2, pp. 113–120, 2019, doi: 10.31294/ijse.v5i2.6963.

[2] N. Armstrong and K. Yoon, "Movie Rating Prediction," *Citeseer*, pp. 507–511, 2009, [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.87.1964&rep=rep1&type=pdf.

[3] W. R. Bristi, Z. Zaman, and N. Sultana, "Predicting IMDb Rating of Movies by Machine Learning Techniques," *2019 10th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2019*, pp. 1–5, 2019, doi: 10.1109/ICCCNT45670.2019.8944604.

[4] R. Dhir and A. Raj, "Movie Success Prediction using Machine Learning Algorithms and their Comparison," *ICSCCC 2018 - 1st Int. Conf. Secur. Cyber Comput. Commun.*, pp. 385–390, 2018, doi: 10.1109/ICSCCC.2018.8703320.

[5] A. G. Sooai and M. Laniwati, "Ekstraksi Ulasan Sentimen Film dari Twitter dengan Naïve Bayes pada Situs Web Media Sosial Penggemar Film," *J. Intell. Syst. Comput.*, vol. 3, no. 1, pp. 49–54, 2021, doi: 10.52985/insyst.v3i1.186.

[6] P. S. M. Suryani, L. Linawati, and K. O. Saputra, "Penggunaan Metode Naïve Bayes Classifier pada Analisis Sentimen Facebook Berbahasa Indonesia," *Maj. Ilm. Teknol. Elektro*, vol. 18, no. 1, p. 145, 2019, doi: 10.24843/mite.2019.v18i01.p22.

[7] P. Antinasari, R. S. Perdana, and M. A. Fauzi, "Analisis Sentimen Tentang Opini Film Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naive Bayes Dengan Perbaikan Kata Tidak Baku," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 12, pp. 1718–1724, 2017, [Online]. Available: http://j-ptiik.ub.ac.id.

[8] R. Yanto, "Implementasi Data Mining Estimasi Ketersediaan Lahan Pembuangan Sampah menggunakan Algoritma Simple Linear Regression," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 2, no. 1, pp. 361–366, 2018, doi: 10.29207/resti.v2i1.282.

[9] M. Alfi, R. Reynaldhi, and Y. Sibaroni, "Analisis Sentimen Review Film pada Twitter menggunakan Metode Klasifikasi Hybrid Naïve Bayes dan Decision Tree," vol. 8, no. 5, pp. 10127–10137, 2021.

[10] J. Prima *et al.*, "PREDIKSI HARGA MOBIL MENGGUNAKAN ALGORITMA REGRESSI DENGAN HYPER-PARAMETER TUNING," vol. 4, no. 2, pp. 1–5, 2021.

[11] J. M. Hellerstein and U. C. Berkeley, "DataCleaning-ucb(2)," *United Nations Econ. Comm. Eur.*, pp. 1–42, 2008, [Online]. Available: http://db.cs.berkeley.edu/jmh/cleaning-unece.pdf%5Cnpapers2://publication/uuid/DC7173AB-6B26-4B8B-AEC3-4C7E65CEEFED.

[12] J. Prima *et al.*, "PREDIKSI WATER QUALITY INDEX (WQI) MENGGUNAKAN ALGORITMA REGRESSI DENGAN HYPER-PARAMETER TUNING," vol. 5, no. 1, pp. 44–50, 2021.

[13] J. Prima *et al.*, "ANALISIS PERBANDINGAN SENTIMEN CORONA VIRUS DISEASE-2019 (COVID19) PADA TWITTER MENGGUNAKAN METODE LOGISTIC REGRESSION DAN SUPPORT VECTOR MACHINE (SVM)," vol. 5, no. 2, 2022.

[14] N. S. Fathullah, Y. A. Sari, and P. P. Adikara, "Analisis Sentimen Terhadap Rating dan Ulasan Film dengan menggunakan Metode Klasifikasi Naïve Bayes dengan Fitur Lexicon-Based," *J.*

Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 4, no. 2, pp. 590–593, 2020, [Online]. Available: http://j-ptiik.ub.ac.id.

[15] A. Saleh, E. Indra, and M. Harahap, "Kombinasi Jaringan Learning Vector Quantization Dan Normalized Cross Correlation Pada Pengenalan Wajah," *J. Sist. Inf. dan Ilmu Komput. Prima(JUSIKOM PRIMA)*, vol. 3, no. 2, pp. 13–20, 2020, doi: 10.34012/jusikom.v3i2.851.

[16] Tamba, S.P., 2022. Penerapan Data Mining Algoritma Apriori Dalam Menentukan Stok Bahan Baku Pada Restoran Nelayan Menggunakan Metode Association Rule. *Jurnal Sistem Informasi dan Ilmu Komputer Prima (JUSIKOM PRIMA)*, *5*(2), pp.97-102.