Journal of Applied Intelligent System (e-ISSN : 2502-9401 | p-ISSN : 2503-0493)
Vol. 8 No. 3, November 2023, pp. 341 – 352
DOI:

Film Review Sentiment Analysis: Comparison of Logistic Regression and Support Vector Classification Performance Based on TF-IDF

Dadan Saepul Ramdan[*1], **Riri Damayanti Apnena**[2]
[1,2]Politeknik TEDC Bandung, Jl. Politeknik - Pasantren KM. 2 Lantai 1, Cibabat, Cimahi Utara, Cibabat, Cimahi Utara, Kota Cimahi, Jawa Barat 40513
*Email : dsramdan@poltektedc.ac.id[*1], riri.damayanti.apnena@poltektedc.ac.id[2]*
**Corresponding author*

Castaka Agus Sugianto[3]
[3]Politeknik TEDC Bandung, Jl. Politeknik - Pasantren KM. 2 Lantai 1, Cibabat, Cimahi Utara, Cibabat, Cimahi Utara, Kota Cimahi, Jawa Barat 40513
Email : castaka@poltektedc.ac.id[3]

Abstract – Film sentiment analysis is a process for evaluating a sentiment value that exists in film reviews, so that positive or negative responses from films can be identified. In this study, a sentiment analysis will be carried out on film reviews on IMBD. The analysis was carried out to find out which reviews were positive and negative from film critics. The method used to carry out sentiment analysis in this study is review analysis and processing with TF-IDF and a positive or negative prediction process based on reviews that have been processed using a logistic regression algorithm and support vector classification. The data to be used is film reviews on IMBD, which consists of 2000 data, which is divided into 1000 positive data and 1000 negative data. Which is where the data will be preprocessed first and split with a percentage of 70% training data and 30% testing data. In the prediction process using the logistic regression algorithm, obtaining a test accuracy of 80.61%. While the prediction process using the support vector classification algorithm obtains a test accuracy of 82.42%.

Keywords – Sentiment Analysis, TF-IDF, Logistic Regression, Support Vector Classification, Film

1. INTRODUCTION

Film is an art form that combines video, sound and narration to be able to convey something to the audience. Because of this, films are a common means of entertainment for everyone [1]. Fithratullah [2], argued that film is considered as part of a work of art that is made based on the needs and desires that emerge from society. Therefore, the popularity of films is usually based on the type of reviews given by the audience [3]. Coupled with current technological advances, it is common that when people watch a film, they give and express their opinions on public social networking sites [4]. Because of this, social media is now a source that can be used to get opinions instantly [5], not only opinions but also someone's statements on topics that are currently trending [6]. So in the analysis of film sentiment, the opinions or opinions given by the audience about the film can be used to find out how the audience feels or responds when watching a particular film. The responses given were divided into two classes, namely positive responses and negative responses. Teixeira et. al [7], argued that film is a process of character formation for events that can occur in a predetermined time and space. Because of this, everyone's preferences and opinions can vary depending on that person's point of view.

Sentiment analysis is a process for processing a text to be able to find out the value or message contained therein [8]. Purnomoputra et. al [9], stated that sentiment analysis can be used to classify films, whether they are good or bad films. The process carried out in sentiment analysis is to do computations to be able to determine expressions or feelings from the reviews given by the audience [10]. Dang et. al [11], stated that usually the data obtained to carry out sentiment analysis is obtained from social media, where the audience provides a lot of information and reviews of something. Sentiment analysis is part of the data mining process related to natural language processing (NLP) [12], which has been a topic of research since early 2000 [13]. In sentiment analysis, the reviews given show how the user responds, responds or reacts to a service or product [14]. So that from the results of the sentiment analysis obtained, it can assist in the decision-making process [15] whether a service is good or bad. In the world of film, the quality of the film itself is obtained from pre-existing audience reviews [16]. So that sentiment analysis has a role to see whether the existing review is a positive or negative review. Therefore, the process of sentiment analysis is included in the data classification method [17] and plays a major role in conducting perspective analysis of the audience about something [18].

Machine learning or often referred to as machine learning (ML) is part of the artificial intelligence family or commonly called artificial intelligence (AI) [19]. In the process, machine learning works with large data to be able to train and optimize models based on algorithms, where these models can later make predictions [20]. In machine learning, we don't need to program the model to do the learning, which means the model can do the learning automatically [21]. TR et. al [22], argued that the machine learning method works by inputting test samples after training and the model learns patterns from existing data. The machine learning method is considered beneficial because the model can learn from mistakes gradually to be able to improve the performance of the model itself by learning more similar data [23]. In this study, the sentiment analysis process will use the TF-IDF algorithm to convert text into vector form and predictions using logistic regression algorithms and support vector classification so that knowledge can be easily extracted from large amounts of data [24]. Machine learning methods can also be useful in many sectors of the economy such as manufacturing, banks, etc. [25], not only in a certain scope.

Term Frequency/Inverse Document Frequency (TF-IDF) is a method used for the mining process of text. TF-IDF is usually used to weigh words based on their uniqueness, so that relevance can be found between words, documents and certain categories. Zhou et. al [26], argued that TF-IDF is a type of measurement in a statistical method that is widely used for data processing in text form. TF-IDF is an effective method for extracting knowledge from an attribute so that the attribute can represent the whole document properly [27]. In the process, calculations are carried out using statistical methods to map or transform text into vectors, then calculate the similarity between the data and the vector text [28].

Logistic regression (LR), is a popular and commonly used algorithm to classify data [29]. Logistic regression is widely used to carry out binary classification processes or classifications that only have 2 class targets [30]. Pan et. al [31], argued that logistic regression is a linear classification method that is easy and simple to use. Logistic regression is included in the type of supervised learning [32]. In the process, logistic regression is used to measure the level of statistical significance of each predictor variable with a probability approach [33].

Support Vector Classification (SVC) is part of the Support Vector Machine (SVM) which has a structured risk minimization principle [34]. The way Support Vector Classification works is the same as the way Support Vector Machine works, namely by minimizing the distance between the decision boundary (Support Vector) and the sample (maximum margin) [35]. So, in the process a hyperplane will be searched for each existing class sample [36]. Djedidi et. al [37], stated that in this method a hyperplane will be sought to be able to divide between positive and negative classes using the most optimal margins.

Research conducted by Soubraylu et. al [38], discussed sentiment analysis based on film reviews using the hybrid convolutional bidirectional recurrent neural network method. This study aims to be able to carry out sentiment analysis and build models using the hybrid deep learning method that combines the convolutional neural network (CNN) method with the bidirectional gated recurrent unit (BGRU) method. The results obtained in this study are that the model built gets better results than other models, namely with an F1-Score of 87.62% and 77.4% with the IMBD and Polarity dataset. In a study conducted by Bodapati et. al [39], discusses sentiment analysis based on film reviews using the Long-Short Term Memory (LSTMs) method. This study aims to be able to build models using the Long-Short Term Memory method or LSTMs to be able to carry out sentiment analysis. The results obtained from this study are that the model built succeeded in obtaining better accuracy compared to other methods, namely 88.46%.

Dalam artikel ini kami telah menginvestigasi proses review analisis untuk film menggunakan TF-IDF berbasis Logistic Regression dan Support Vector Classification. Dari beberapa penelitian yang telah dilkukan, terdapat salah satu algortima saja yang digunakan missal Logistic Regression saja atau Support Vector Classification saja. Akurasi dari masing-maisng algrotima masih dapat ditingkatkan dengan menambah parameter seperti yang kami lakukan dan telah dijelaskan pada sub bab berikutnya. Diketahui bahwa Logistic Regression dan Support Vector Classification dpat

2. RESEARCH METHOD

2.1. Dataset

Film is an art form that combines video, sound and narration to be able to convey something to the audience. Teixeira et. al [7], argued that film is a process of character formation for events that can occur in a predetermined time and space. In this study, a sentiment analysis of film reviews will be carried out. Sentiment analysis is part of the data mining process Sentiment analysis can be used to classify films, whether they are good or bad films [9], therefore the sentiment analysis process is related to natural language processing or often called Natural Language Processing (NLP) [11]. So that from the results of the sentiment analysis obtained, it can help in the decision-making process [15] whether to watch the film or not. Because the quality of the film itself is obtained from the reviews of pre-existing audiences [16]. In the research that will be conducted, use the Pang and Lee's Movie Review Data dataset obtained from the link http://boston.lti.cs.cmu.edu/classes/95-865-K/HW/HW3/. The data to be used amounts to 2000 data with 2 columns. The first column is used as the target variable and the second column is used as the predictor variable. Of the 2000 existing data, it is divided into 2 main classes, namely positive and negative. With the amount of data for each class, namely 1000 positive data and 1000 negative data. For the predictor column, the data type used is string, which contains movie reviews. The purpose of using this dataset is to be able to build a model that can carry out sentiment analysis so that it can distinguish between positive and negative reviews from the existing data.

Of the 2000 total existing data, it will be further divided into 2 data, namely training data and testing data. The division is done by calculating the percentage of data as much as 70% training data and 30% testing data. With target data for each data, both training data and test data have as many as 2 classes, namely the positive class and the negative class. The purpose of sharing data is to carry out development and testing on models with predetermined algorithms. The training data is used to train the model using the algorithm used. In this study, the algorithm used is a logistic regression classifier and also a support vector classifier. After training on the model, model testing will be carried out using test data, so that performance measurements can be carried out from the results of the sentiment analysis model training.

2.2. Term Frecuency/Inverse Document Frecuency (TF-IDF)

Term Frequency/Inverse Document Frequency or can be abbreviated as TF-IDF is a method that is usually used for processes related to natural language processing or often called Natural Language Processing (NLP). TF or Term Frequency is a process for comparing a word that appears in the text with the total number of words in the text. For TF calculations can be seen in (1). IDF or Inverse Document Frequency is a process to measure how unique a word from a corpus or group of words is in the text, so that for DF calculations, corpus data will be searched that contains text or reviews. For IDF calculations, see (2). To get the TF-IDF score, it will be multiplied between the TF value and the IDF score, for the calculation can be seen in (3).

$$TF(Word, Document\ ) = \frac{\sum(the\ word\ appears\ in\ document)}{\sum(word\ in\ document)} \tag{1}$$

$$IDF(Word, Corpus) = \log(\frac{N}{(1 + DF(Word, Corpus))}) \tag{2}$$

$$TF - IDF(Word, Document, Corpus) = TF(Word, Document\ ) * IDF(ord, Corpus) \tag{3}$$

In this study, along with the TF-IDF process, text data will also be processed into vectors so that the resulting text data is in the form of a TF-IDF matrix. Which later the data will be used as training material and also testing of the model.

2.3. Logistic regression (LR)

Logistic regression is widely used to carry out binary classification processes or classifications that only have 2 class targets [30]. Pan et. al [31], argued that logistic regression is a linear classification method that is easy and simple to use. Logistic regression is included in the type of supervised learning [32]. The process carried out in logistic regression is to carry out a linear transformation from features to probability values by using a logistic function or it can be called a sigmoid function. So that the output issued produces a value of 0 or 1. The mathematical formula for logistic regression can be seen in (4).

$$P(C = 1 \mid Z) = \frac{1}{1 + e^{-(q_0 + q_1 * Z_1 + q_2 * Z_2 + \cdots + q_n * Z_n)}} \tag{4}$$

Where :

- $P(C = 1 \mid Z)$ is the probability in class 1 with input Z
- e is an Euler number, namely 2.71828
- q_0 , q_1, q_2, q_n is the model parameter used during training
- Z_1, Z_2, Z_n is a predictor variable or feature

2.4. Support Vector Classification (SVC)

Support Vector Classification (SVC) is part of the Support Vector Machine (SVM) which is usually used to classify data and has a structured risk minimization principle [34]. Support Vector Classification works by building a hyperplane which aims to divide data into 2 classes according to existing targets. Djedidi et. al [37], stated that in this method a hyperplane will be sought to be able to divide between positive and negative classes using the most optimal margins. So that the data around the hyperplane is called a support vector whose job is to help determine the position of the hyperplane. The SVC mathematical formula can be seen in (5).

$$\omega * Z + b = 0 \tag{5}$$

Where :

- ω is the weight vector of the hyperplane
- X is the feature used
- b is the value of bias (shift)

2.5. Confusion Matrix

Performance calculations are carried out after completing the model testing process. This performance calculation aims to see how the robustness of the model is when dealing with new data other than the data used during training. Performance calculations can be done in many ways. In this study, the calculation of model performance is carried out by looking at the accuracy of the test being sought by comparing the model's predicted results with the actual results, the confusion matrix of the test and also the test classification report from the Logistic Regression or Support Vector Classification model that has been built.

Confusion Matrix is the result of a calculation that compares the predicted value of the model with the original or actual value. There are 4 values that are calculated in the confusion matrix, namely the TP value (true positive) or the positive value that was successfully predicted correctly, the TN value (true negative) or the negative value that was predicted correctly, the FP value (false positive) or the negative value that was incorrectly predicted to be positive and FN values (false negatives) or positive values that are incorrectly predicted to be negative.

2.6. Classification Report

After obtaining the values in the confusion matrix, the values in the classification report can be calculated. The values to be calculated in the classification report are precision, recall, f1-score and support values. Where precision is the result of a calculation that shows what percentage of positive predictions are correct from the model. For precision calculations, see (6). Recall is a calculation of the model's performance in identifying all positive cases in the data. For recall calculations, it can be seen in (7) F1-Score is the calculated value of the harmonic average between precision and recall values, so that a balance can be achieved between precision and recall values. For the calculation of the f1-score, see (8). While support is the amount of data used to perform confusion matrix calculations.

$$P = \frac{TruePos}{(TruePos + FalsePos)} \tag{6}$$

$$R = \frac{TruePos}{(TruePos + FalseNeg)} \tag{7}$$

$$f1 = 2 * \frac{(P * R)}{(P + R)} \tag{8}$$

2.7. Proposed Scheme

In the process of conducting sentiment analysis based on film reviews, this study will use a jupyter notebook as an IDE to write programs and the Python programming language to be able to implement the system. The first step is to prepare the data to be used either for training or testing. The data that will be used in this research is Pang and Lee's Movie Review data obtained from the link http://boston.lti.cs.cmu.edu/classes/95-865-K/HW/HW3/. After obtaining the data, the data can be read and processed, then trained and tested using logistic regression (LR) and support vector classification (SVC) algorithms. So that after training and testing, the performance of the model from the test results can be calculated.

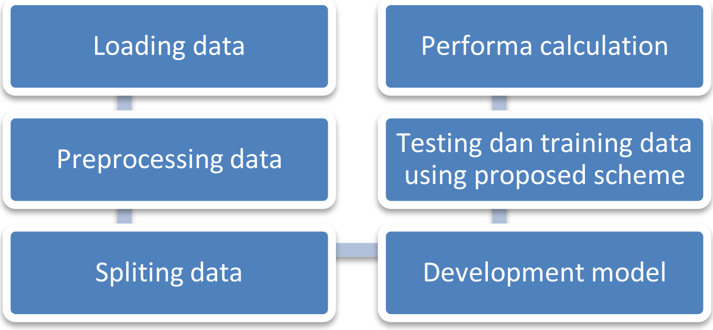


Figure 1. Our Sentiment Analysis Scheme

For an illustration of the workflow, it can be seen in Figure (1) and an explanation of the process, which can be seen in the description below Figure (1), as shown in stages as follows :

1. Read data with the .csv extension that has been prepared
2. After the data is read, data preprocessing will be carried out.

- Looking for a corpus or a collection of words that often appear in the dataset according to the existing class.

```
for val in data[data[x] == y].text:
 text = val.lower()
 tokens = nltk.word_tokenize(text)
 for words in tokens:
 wordsX = wordsX + words + ' '
```

- Changing data labels that were originally categorical to be numerical.

```
 data = data.replace(['x','y'],[0, 1])
```

- Clears every text in the data by removing conjunctions.

```
  text = text.translate(str.maketrans('', '', string.punctuation))
  text = [word for word in text.split() if word.lower() not in stopwords.words('english')]
```

- Converting text data into vectors using the TF-IDF method, so that the data can be used as training materials and model testing. In this research, we will use help from the sklearn python library to call the *TfidfVectorizer()* function.

```
        vectorizer = TfidfVectorizer()
        vectors = vectorizer.fit_transform(data['x'])
        features = vectors
```

3. After data preprocessing, the processed data is divided into training data and testing data, with a percentage of 70% training data and 30% testing data.
4. Then after the data is divided into training data and testing data, machine learning models will be built using Logistic Regression or Support Vector Classification algorithms. Which is where this model will later be used to conduct training and testing based on data that has been preprocessed. For the parameters used in each model can be seen in Table (1) and (2).

Table 1. Parameter Logistic Regression

penalty	l2
dual	False
tol	0.0
C	1.0
fit_intercept	True
intercept_scaling	1
class_weight	None
random_state	None
solver	sag
max_iter	100
multi_class	auto
verbose	0
warm_start	False
n_jobs	None

Table 2. Parameter Support Vector Classification

C	1.0
kernel	sigmoid
degree	3
gamma	1.0
coef0	0.0
shrinking	True
probability	False
tol	0.001
class_weight	balanced
verbose	False
max_iter	-1
decision_function_shape	ovr
random_state	None

5. After the model is built along with the specified parameters, then the model training process is carried out using the training data. Then, after the model has been trained, the model will be tested using test data, which aims to see how the model performs when given new data.
6. Then after the training and testing process has been completed, the results of the testing process can be calculated to produce a classification metric or classification report and confusion matrix.

3. RESULTS AND DISCUSSION

This study uses the Jupyter Notebook IDE to write program code and the Python programming language to implement a sentiment analysis system based on film reviews. For the data used, it consists of a total of 2000 data where the data is divided into 2 classes, namely the positive class and the negative class. With the amount of data in each class, namely 1000 positive class data and 1000 negative class data. After data preparation, the data will be preprocessed before the data is used as training materials and model testing. Preprocessing is done by looking for corpus data or data that often appears, then changing the labels that were initially categorical to numerical. After that, the data whose labels have been transformed are processed again by removing connecting words or stop words from the entire data. After the data is processed in such a way, the data which no longer has stop words is converted into a vector using the TF-IDF method, so that the text data that has become a vector will be used as training material and model testing.

After the data has been preprocessed, a logistic regression model or support vector classification will be built which later this model will be used to be able to predict whether a given review will be a positive or negative review. After the model is finished to be built along with its parameters, then training is carried out on the model so that the model can recognize patterns from the data provided. Then after the training process is complete, the model testing process is carried out again which aims to evaluate the model and perform model performance calculations. After testing the model using logistic regression, the accuracy of the test is 80.61%. For the confusion matrix testing with logistic regression can be seen in Figure (2). Meanwhile, after testing the model with support vector classification, the accuracy of the model test is 82.42%. For the confusion matrix testing with support vector classification can be seen in figure (3).

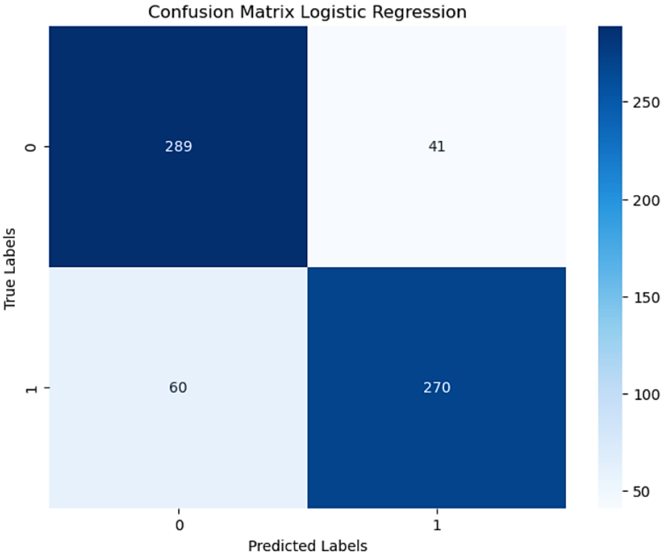


Figure 2. Confusion Matrix of Logistic Regression

It can be seen in Figure (2) and Figure (3) that shows the results of the confusion matrix from the model testing that has been done. In the confusion matrix, label 0 indicates the positive class and label 1 indicates the negative class. The results obtained from testing the logistic regression model are TP values of 289 data, FN values of 60 data, TN values of 270 data and FP values of 41. Meanwhile, in testing using support vector classification, TP values of 268 data, FN values of 54 data, the TN value is 276 data and the FP value is 62 data. After knowing the true positive, true negative, false positive and false negative values of the confusion matrix, these values can be used to calculate precision (5), recall (6), f1-score (7) model test results, both with the logistic regression algorithm or a support vector classification algorithm. For precision, recall, f1-score calculation results from the logistic regression model and support vector classification can be seen in figures (4) and (5).

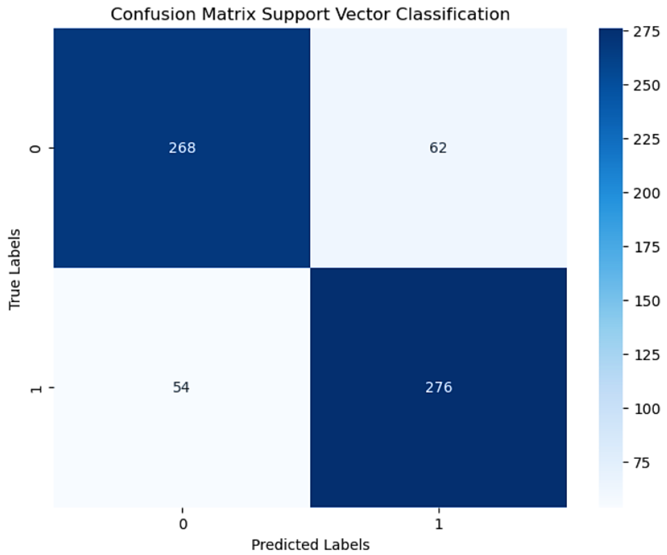
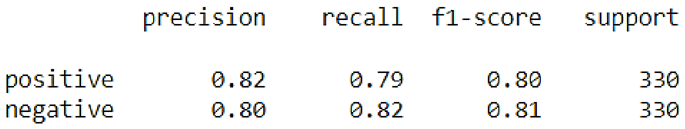


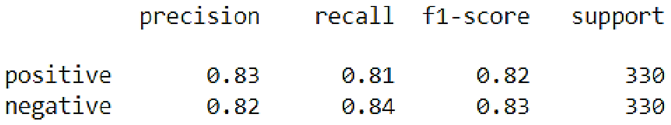
Figure 3. Confusion Matrix of Support Vector Classification



```
              precision    recall  f1-score   support

positive           0.82      0.79      0.80       330
negative           0.80      0.82      0.81       330
```

Figure 4. Classification Report for Logistic Regression



```
              precision    recall  f1-score   support

positive           0.83      0.81      0.82       330
negative           0.82      0.84      0.83       330
```

Figure 5. Classification Report for Support Vector Classification

Figures (4) and (5), show the results of the classification report from the process using the logistic regression algorithm and also the support vector classification. From the results of these two processes, it can be seen that the best precision value is obtained in the positive class using the support vector classification algorithm, which is 83%. The best recall value is obtained from the negative class with the support vector classification algorithm. Meanwhile, the best f1-score is obtained in the negative class using the support vector classification algorithm. With this description, it can be seen that using the support vector classification can make a good guess at the positive class, can identify all the positive cases in the negative class and also has a better balance of harmonic values or values between precision and recall, compared to a logistic regression algorithm.

4. CONCLUSION

In the tests that have been carried out using the logistic regression method and support vector classification to be able to carry out sentiment analysis based on film reviews, the aim is for the model to be able to classify the reviews that are given including positive or negative reviews, get the result that the test uses the logistic regression method get a test accuracy of 80.61%, while using the support vector classification method get an accuracy of 82.42%. From the test results that have been obtained, it can be concluded that using the support vector classification method is more effective and accurate in being able to carry out sentiment analysis based on film reviews.

It is expected to be able to add the testing methods used, such as using random forests for next research, decision trees or methods related to neural networks such as CNN (Convolutional Neural Network), RNN (Recurrent Neural Network), LSTM (Long-Short Term Memory), and so on to see how the performance and effectiveness of other methods are. And it is also hoped that in future research it can add parameters or manipulate parameters in logistic regression or support vector classification so that the model training process can be more complex and expected to be more accurate.

REFERENCES

[1] Astuti, R. W., Waluyo, H. J., & Rohmadi, M. (2019). Character Education Values in Animation Movie of Nussa and Rarra. Budapest International Research and Critics Institute (BIRCI-Journal) : Humanities and Social Sciences, 2(4), 215–219. https://doi.org/10.33258/birci.v2i4.610

[2] Fithratullah, M. (2021). Representation of Korean values sustainability in American remake movies. Teknosastik, 19(1), 60-73. https://doi.org/10.33365/ts.v19i1.874

[3] Pavitha, N., Pungliya, V., Raut, A., Bhonsle, R., Purohit, A., Patel, A., & Shashidhar, R. (2022). Movie recommendation and sentiment analysis using machine learning. Global Transitions Proceedings, 3(1), 279-284. https://doi.org/10.1016/j.gltp.2022.03.012

[4] Rahman, A., & Hossen, M. S. (2019, September 1). Sentiment Analysis on Movie Review Data Using Machine Learning Approach. 2019 International Conference on Bangla Speech and Language Processing, ICBSLP 2019. https://doi.org/10.1109/ICBSLP47725.2019.201470

[5] Rehman, A. U., Malik, A. K., Raza, B., & Ali, W. (2019). A Hybrid CNN-LSTM Model for Improving Accuracy of Movie Reviews Sentiment Analysis. Multimedia Tools and Applications, 78(18), 26597–26613. https://doi.org/10.1007/s11042-019-07788-7

[6] A. M. Rahat, A. Kahir and A. K. M. Masum, "Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset," 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART), Moradabad, India, 2019, pp. 266-270, doi: 10.1109/SMART46866.2019.9117512.

[7] Teixeira, M. B. M., Galvão, L. L. da C., Mota-Santos, C. M., & Carmo, L. J. O. (2021). Women and work: film analysis of Most Beautiful Thing. In Revista de Gestao (Vol. 28, Issue 1, pp. 66–83). Emerald Group Holdings Ltd. https://doi.org/10.1108/REGE-03-2020-0015

[8] Kumar, K., Harish, B. S., & Darshan, H. K. (2019). Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method. International Journal of Interactive Multimedia and Artificial Intelligence, 5(5), 109. https://doi.org/10.9781/ijimai.2018.12.005

[9] Bintang Purnomoputra, R., & Novia Wisesty, U. (2019). Sentiment Analysis of Movie Reviews using Naïve Bayes Method with Gini Index Feature Selection. OPEN ACCESS J DATA SCI APPL, 2(2), 85–094. https://doi.org/10.34818/JDSA.2019.2.36

[10] Kumar, S., De, K., & Roy, P. P. (2020). Movie Recommendation System Using Sentiment Analysis from Microblogging Data. IEEE Transactions on Computational Social Systems, 7(4), 915–923. https://doi.org/10.1109/TCSS.2020.2993585

[11] Dang, N. C., Moreno-García, M. N., & de la Prieta, F. (2020). Sentiment analysis based on deep learning: A comparative study. Electronics (Switzerland), 9(3). https://doi.org/10.3390/electronics9030483

[12] Bonta, V., Kumaresh, N., & Janardhan, N. (2019). A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis. Asian Journal of Computer Science and Technology, 8(S2), 1–6. https://doi.org/10.51983/ajcst-2019.8.s2.2037

[13] Behera, R. K., Jena, M., Rath, S. K., & Misra, S. (2021). Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data. *Information Processing and Management*, *58*(1). https://doi.org/10.1016/j.ipm.2020.102435

[14] Li, L., Goh, T. T., & Jin, D. (2020). How textual quality of online reviews affect classification performance: a case of deep learning sentiment analysis. Neural Computing and Applications, 32(9), 4387–4415. https://doi.org/10.1007/s00521-018-3865-7

[15] Malviya, S., Tiwari, A. K., Srivastava, R., & Tiwari, V. K. (2020). Machine Learning Techniques for Sentiment Analysis: A Review. SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology, 12(2), 72–78. https://doi.org/10.18090/samriddhi.v12i02.3

[16] Maulana, R., Rahayuningsih, P. A., Irmayani, W., Saputra, D., & Jayanti, W. E. (2020). Improved Accuracy of Sentiment Analysis Movie Review Using Support Vector Machine Based Information Gain. Journal of Physics: Conference Series, 1641(1). https://doi.org/10.1088/1742-6596/1641/1/012060

[17] Qaisar, S. M. (2020, October 13). Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory. 2020 2nd International Conference on Computer and Information Sciences, ICCIS 2020. https://doi.org/10.1109/ICCIS49240.2020.9257657

[18] Haque, M. R., Akter Lima, S., & Mishu, S. Z. (2019). Performance Analysis of Different Neural Networks for Sentiment Analysis on IMDb Movie Reviews. 3rd International Conference on Electrical, Computer and Telecommunication Engineering, ICECTE 2019, 161–164. https://doi.org/10.1109/ICECTE48615.2019.9303573

[19] Sharma, N., Sharma, R., & Jindal, N. (2021). Machine Learning and Deep Learning Applications-A Vision. Global Transitions Proceedings, 2(1), 24–28. https://doi.org/10.1016/j.gltp.2021.01.004

[20] Wei, J., Chu, X., Sun, X. Y., Xu, K., Deng, H. X., Chen, J., Wei, Z., & Lei, M. (2019). Machine learning in materials science. In InfoMat (Vol. 1, Issue 3, pp. 338–358). Blackwell Publishing Ltd. https://doi.org/10.1002/inf2.12028

[21] Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised Classification Algorithms in Machine Learning: A Survey and Review. Advances in Intelligent Systems and Computing, 937, 99–111. https://doi.org/10.1007/978-981-13-7403-6_11

[22] Ramesh, T. R., Lilhore, U. K., Poongodi, M., Simaiya, S., Kaur, A., & Hamdi, M. (2022). PREDICTIVE ANALYSIS OF HEART DISEASES WITH MACHINE LEARNING APPROACHES. Malaysian Journal of Computer Science, 2022(Special Issue 1), 132–148. https://doi.org/10.22452/mjcs.sp2022no1.10

[23] Lee, C. S., & Lee, A. Y. (2020). Clinical applications of continual learning machine learning. In The Lancet Digital Health (Vol. 2, Issue 6, pp. e279–e281). Elsevier Ltd. https://doi.org/10.1016/S2589-7500(20)30102-3

[24] Breck, E., Polyzotis, N., Roy, S., Whang, S., & Zinkevich, M. (2019, April). Data Validation for Machine Learning. In MLSys. Proceedings of the 2 nd SysML Conference, Palo Alto, CA, USA, 2019

[25] Huy, D. T. N., Le, T. H., Hang, N. T., Gwoździewicz, S., Trung, N. D., & Van Tuan, P. (2021). Further researches and discussion on machine learning meanings-and methods of classifying and recognizing users gender on internet. Advances in Mechanics, 9(3), 1190-1204.

[26] Zhou, Z., Qin, J., Xiang, X., Tan, Y., Liu, Q., & Xiong, N. N. (2020). News text topic clustering optimized method based on TF-iDF algorithm on spark. Computers, Materials and Continua, 62(1), 217–231. https://doi.org/10.32604/cmc.2020.06431

[27] Dalaorao, G. A., Sison, A. M., & Medina, R. P. (2019). Integrating Collocation as TF-IDF Enhancement to Improve Classification Accuracy. 2019 IEEE 13th International Conference on Telecommunication Systems, Services, and Applications (TSSA). doi:10.1109/tssa48701.2019.8985458

[28] Wang, J., Xu, W., Yan, W., & Li, C. (2019). Text similarity calculation method based on hybrid model of LDA and TF-IDF. ACM International Conference Proceeding Series, 1–8. https://doi.org/10.1145/3374587.3374590

[29] Samsudin, N. M., Mohd Foozy, C. F. B., Alias, N., Shamala, P., Othman, N. F., & Wan Din, W. I. S. (2019). Youtube spam detection framework using naïve bayes and logistic regression. Indonesian Journal of Electrical Engineering and Computer Science, 14(3), 1508–1517. https://doi.org/10.11591/ijeecs.v14.i3.pp1508-1517

[30] Zou, X., Hu, Y., Tian, Z., & Shen, K. (2019). Logistic Regression Model Optimization and Case Analysis. *Proceedings of IEEE 7th International Conference on Computer Science and Network Technology, ICCSNT 2019*. https://doi.org/10.1109/ICCSNT47585.2019.8962457

[31] Luo, H., Pan, X., Wang, Q., Ye, S., & Qian, Y. (2019). Logistic regression and random forest for effective imbalanced classification. Proceedings - International Computer Software and Applications Conference, 1, 916–917. https://doi.org/10.1109/COMPSAC.2019.00139

[32] Alotaibi, F. M. (2019). Classifying text-based emotions using logistic regression. http://dx.doi.org/10.21015/vtcs.v16i2.551

[33] Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. Augmented Human Research, 5(1). https://doi.org/10.1007/s41133-020-00032-0

[34] Robles-Velasco, A., Cortés, P., Muñuzuri, J., & Onieva, L. (2020). Prediction of pipe failures in water supply networks using logistic regression and support vector classification. Reliability Engineering and System Safety, 196. https://doi.org/10.1016/j.ress.2019.106754

[35] Rákos, O., Aradi, S., & Bécsi, T. (2020). Lane change prediction using Gaussian classification, support vector classification and neural network classifiers. Periodica Polytechnica Transportation Engineering, 48(4), 327–333. https://doi.org/10.3311/PPTR.15849

[36] Liu, W., & Rao, Z. (2020). Road Icing Warning System Based on Support Vector Classification. IOP Conference Series: Earth and Environmental Science, 440(5). https://doi.org/10.1088/1755-1315/440/5/052071

[37] Djedidi, O., Djeziri, M. A., Morati, N., Seguin, J. L., Bendahan, M., & Contaret, T. (2021). Accurate detection and discrimination of pollutant gases using a temperature modulated MOX sensor combined with feature extraction and support vector classification. Sensors and Actuators, B: Chemical, 339. https://doi.org/10.1016/j.snb.2021.129817

[38] Soubraylu, S., & Rajalakshmi, R. (2021). Hybrid convolutional bidirectional recurrent neural network based sentiment analysis on movie reviews. Computational Intelligence, 37(2), 735–757. https://doi.org/10.1111/coin.12400

[39] Bodapati, J. D., Veeranjaneyulu, N., & Shaik, S. (2019). Sentiment analysis from movie reviews using LSTMs. Ingenierie Des Systemes d'Information, 24(1), 125–129. https://doi.org/10.18280/isi.240119