Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, January 2022
DOI : https://doi.org/10.33395/sinkron.v7i1.11204
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Analysis Sentiment based on IMDB aspects from movie reviews using SVM

Nur Ghaniaviyanto Ramadhan[1)*], Teguh Ikhlas Ramadhan[2)]
[1)] Institut Teknologi Telkom Purwokerto, Indonesia, [2)] Universitas Perjuangan Tasikmalaya [1)]ghani@ittelkom-pwt.ac.id, [2)] teguhikhlas@unper.ac.id

Submitted : Nov 2. 2021 | **Accepted** : Jan 1, 2022 | **Published** : Jan 10, 2022

Abstract: A movie is a spectacle that can be done at a relaxed time. Currently, there are many movies that can be watched via the internet or cinema. Movies that are watched on the internet are sometimes charged to watch so that potential viewers before watching a movie will read comments from users who have watched the movie. The website that is often used to view movie comments today is IMDB. Movie comments are many and varied on the IMDB website, we can see comments based on the star rating aspect. This causes users to have difficulty analyzing other users' comments. So, this study aims to analyze the sentiment of opinions from several comments from IMDB website users using the star rating aspect and will be classified using the support vector machine method (SVM). Sentiment analysis is a classification process to understand the opinions, interactions, and emotions of a document or text. SVM is very efficient for many applications in science and engineering, especially for classification (pattern recognition) problems. In addition to the SVM method, the TF-IDF technique is also used to change the shape of the document into several words. The results obtained by applying the SVM model are 79% accuracy, 75% precision, and 87% recall. The SVM classification is also superior to other methods, namely logistic regression.

Keywords: Analysis Sentiment, Movie Reviews, IMDB, Support Vector Machine, TF-IDF

INTRODUCTION

A movie is a spectacle that can be done at a relaxed time. Currently, there are many movies that can be watched via the internet or cinema. Movies that are watched on the internet are sometimes charged to watch so that potential viewers before watching a movie will read comments from users who have watched the movie. The website that is often used to view movie comments today is IMDB. On this website, one can search for the film they want to watch by reading the comments first to determine which film to watch based on the most positive or negative comments. Movie comments are many and varied on the IMDB website, you can see comments based on the star rating aspect. This causes users to have difficulty analyzing other users' comments.

Sentiment analysis is a suitable process in the above problems (Tripathi, A., & Trivedi, S. K. 2016). Sentiment analysis is a classification process to understand the opinions, interactions, and emotions of a document or text (Zamzami, F. N., & Adiwijaya, A. 2021), (Shivaprasad, T. K., & Shetty, J. 2017), (Zainuddin, N., & Selamat, A. 2014). Currently, there are many studies related to sentiment analysis problems related to movie reviews. For example, in paper (Tripathi, A., & Trivedi, S. K. 2016) explains that now it is very easy to get feedback on a product or film. The author (Tripathi, A., & Trivedi, S. K. 2016) explains that feature selection in sentiment analysis can be used to help increase the accuracy of sentiment classification. This research (Tripathi, A., & Trivedi, S. K. 2016) uses data from IMDB film India. Research (Mahyarani, M., Adiwijaya, A. 2021) discusses where films should be classified based on sentiment. Paper (Mahyarani, M., Adiwijaya, A. 2021) uses the Naïve Bayes method for classification, TF-IDF as feature extraction, and Information Gain as feature selection with datasets derived from IMDB. Paper (Abidin, Z., Destian, W. 2021) aims to classify the sentiment analysis of film reviews obtained from the IMDb website. The support vector machine (SVM) method is used to classify film review sentiments. Meanwhile, kernel radial basis function (RBF) and information gain (IG) are used to improve classification (Abidin, Z., Destian, W. 2021). Study (Ding, Z., Qi, Y., & Lin, D.2021) used Albert's model to create classifiers and used a "movie review dataset" issued by Stanford University for network training.

*name of corresponding author

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

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, January 2022
DOI : https://doi.org/10.33395/sinkron.v7i1.11204
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Based on the explanation of several previous studies, this study aims to analyse the sentiment of opinions from several comments from IMDB website users using the star rating aspect and will be classified using the support vector machine method.

LITERATURE REVIEW

Paper (Dashtipour, K., Gogate, M. 2021) presents a new, context-aware, context-aware approach to Persian sentiment analysis, driven by deep learning. In particular, the proposed deep learning-driven automated feature engineering approach classifies Persian film reviews as having positive or negative sentiments. Two deep learning algorithms are applied, convolutional neural networks (CNN) and Long Short-Term Memory (LSTM) (Dashtipour, K., Gogate, M. 2021). In paper (Asriyanti, I. P., & Adiwijaya, A. I. 2018) explains that there are data problems contained in the current movie review which causes sentiment analysis to be very slow and less sensitive. In this study, the author performs feature selection to choose to use Information Gain to eliminate features that are not important. Research (Asriyanti, I. P., & Adiwijaya, A. I. 2018) uses dataset V2.0 from Cornell University which contains 1000 positive data and 1000 negative data. In paper (Ulfa, M. A., Irmawati, B. 2018) raised the issue of comments related to Lombok tourism on Twitter social media. The data used in this study amounted to 500 data and used English. Research (Ulfa, M. A., Irmawati, B. 2018) aims to build a sentiment analysis system using the Naïve Bayes classification method, and Mutual Information. In the study (Ulfa, M. A., Irmawati, B. 2018) the Mutual Information feature selection method was able to reduce the classification time by 51.52% and increase the accuracy by 1.7%.

Paper (Gandhi, U. D., Kumar, P. M. 2021) identifies each tweet word and we assign meaning to it. The working features are combined with tweet words, words2vec, stop words and integrated into deep learning technique of Convolution neural network model and Short Long-Term Memory, this algorithm can identify stop word count pattern with its strategy. Both models were well trained and applied to the IMDB dataset containing 50,000 film reviews (Gandhi, U. D., Kumar, P. M. 2021). The study (Rahman, A., & Hossen, M. S. 2019) discusses data from various sources such as medical, social media, newspapers, and film reviews that can be used in data analysis. The author (Rahman, A., & Hossen, M. S. 2019) has collected film review data and used five types of machine learning classifiers to analyze this data. Therefore, the classifiers considered are Bernoulli Naïve Bayes (BNB), Decision Tree (DE), Support Vector Machine (SVM), Maximum Entropy (ME), and Multinomial Naïve Bayes (MNB) (Rahman, A., & Hossen, M. S. 2019). In paper (Mumtaz, D., & Ahuja, B. 2016) proposed a Senti-lexical algorithm to find the polarity of the review as positive, negative, or neutral. The author has also proposed a method for dealing with words that hurt the review and the role of emoticons is also discussed (Mumtaz, D., & Ahuja, B. 2016).

Study [14] proposed different approaches in extracting text features such as bag-of-words model, using large film review corpus, limiting adjectives and adverbs, handling negation, limiting word frequency with thresholds, and using WordNet's synonym knowledge. The authors (Yessenov, K., & Misailovic, S. 2009) evaluated their effect on the accuracy of four machine learning methods – Naive Bayes, Decision Tree, Maximum Entropy, and clustered K-Means. This paper (Lee, S. H., Cui, J. 2016) aims to suggest a way to build a dictionary that is adapted to reflect the characteristics of the data domain. Notably, in the study (Lee, S. H., Cui, J. 2016), the film review data were divided by genre and built a genre-adjusted dictionary. Adjusted dictionary performance in sentiment analysis compared to general sentiment dictionaries. In the study (Lee, S. H., Cui, J. 2016), IMDb data were selected as the subject of analysis, and film reviews were categorized by genre. Six genres on IMDb, 'action', 'animation', 'comedy', 'drama', 'horror' and 'sci-fi' are selected.

METHOD

Fig. 1 is a diagram of the proposed system in this study.

*name of corresponding author

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

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, January 2022
DOI : https://doi.org/10.33395/sinkron.v7i1.11204
e-ISSN : 2541-2019
p-ISSN : 2541-044X

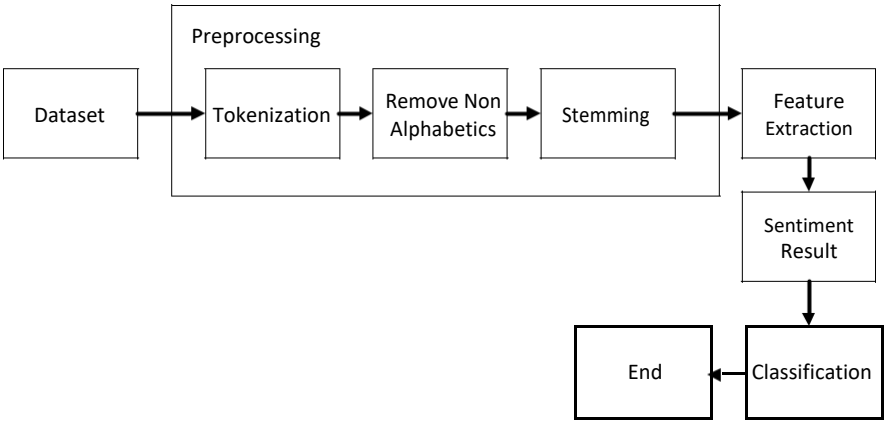

Fig. 1 System Diagram Proposed

Dataset

The dataset was obtained from the IMDB website (IMDB Movie. 2021). This study uses a movie review dataset from the squid game film. The dataset retrieval technique uses a random sampling technique. Random sampling is used at times when processing the entire dataset is not required and is considered too expensive in terms of response time or resource usage (Olken, F., & Rotem, D. 1986). Table 1 shows the characteristics of the dataset used in this study.

Table 1
Characteristics Dataset

No	Features	Type
1	Tag Review	Text
2	Comment Review	Text
3	Rating	Numeric

In table 2 is an example of a dataset.

Table 2
Example Dataset

Tag Review	Comment Review	Rating
Amazing show	This was a show that I felt loved up to the hype. The concept was a brutal, with the end result being so heartbreaking. My only gripe was that by the last few episodes I was starting to tune out a bit. Other than that, this show gets a big recommendation.	8
I don't get this	I really don't know what all the fuss is about... The acting is strong (as expected from Korean actors and actresses) but most people are talking about the plot, which to me, is nothing special. Sure there are some plot twists but there are also way too many gaps for this to be called a "classic" or "masterpiece".	6
Lame	Acting 3/10. Story 1/10. Visuals 6/10. Feels like a B rated Hollywood movie from the 80s. Is it reddit or 4chan behind the hype again?	2
Overhyped and boring.	I don't know why people hype this up so much. Story is boring and predictable. The VIPs were bad actors, and the other characters were mostly annoying. As said in my first sentence I can't seem to understand why people like this so much and that it is rated an 8.3..	4

Tokenization

This process is carried out by removing symbols and characters in sentences, separating sentences into several words, and changing the form of uppercase letters to lowercase letters. Table 3 is an example of the implementation of tokenization.

Table 3

*name of corresponding author

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

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, January 2022
DOI : https://doi.org/10.33395/sinkron.v7i1.11204
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Dataset Tokenization

Tag Review	Comment Review
“amazing”, “show”	“this”, “was”, “a”, “show”, “that”, “i”, “felt”, “loved”, “up”, “to”, “the”, “hype”, “the”, “concept”, “was”, “a”, “brutal”, “with”, “the”, “end”, “result”, “being”, “so” “heartbreaking”, “my”, “only”, “gripe”. “was”, “that”, “by”, “the”, “last”, “few”, “episodes”, “i”, “was”, “starting”, “to”, “tune”, “out”, “a”, “bit”, “other”, “than”, “that”, “this”, “show”, “gets”, “a”, “big”, “recommendation”

Stop Word Removal

In this process, stop words are carried out based on several words that have been determined as in figure 2. The purpose of this process is used to determine the important words used as filtering where the words are from the tokenization process.

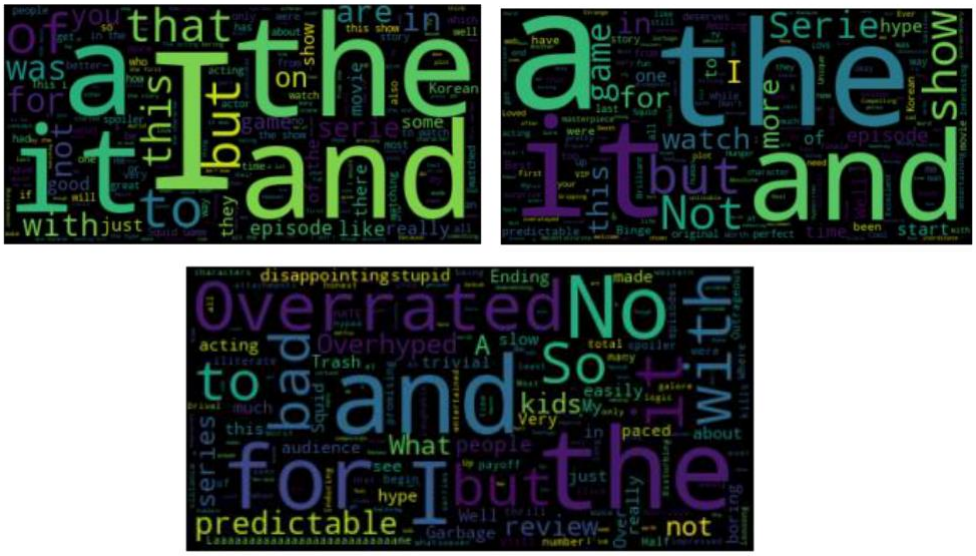

Fig. 2 Stop Word Removal

Stemming

Stemming is a process without variations of the word form into a representative general form (Kannan, S., Gurusamy, V. 2014). For example, the form of the word "acting" is changed to the form "act".

Feature Extraction

This process is generated in the form of features from the movie review document. The technique used is TF-IDF. Term Frequency Inverse Document Frequency (TF-IDF) is one method to calculate the weighting after the feature extraction process (Zamzami, F. N., & Adiwijaya, A. 2021). Weighting using TF-IDF is done by measuring how many words appear in the document (Zainuddin, N., & Selamat, A. 2014). The TF-IDF formula is as follows
(1).

$$(\quad , \quad) = \quad (\quad , \quad) * \log (\qquad \frac{\qquad}{\qquad} \qquad) \tag{1}$$

Where the value of *N* represents the number of all documents, *tk* is the kth word of the keyword, *dj* is the *j-th* document, and *dft* represents the term containing.

Sentiment Result

At this stage, conditions are added based on the rating aspect. If the rating is greater than 5 then it is given a positive value, otherwise, if the rating is less than 5 then it is given a negative value. So at this stage, a new feature is generated called sentiment.

*name of corresponding author

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

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, January 2022
DOI : https://doi.org/10.33395/sinkron.v7i1.11204
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 4
Sentiment Result

Tag Review	Comment Review	Aspect Rating	Sentiment
Amazing show	This was a show that I felt loved up to the hype. The concept was a brutal, with the end result being so heartbreaking. My only gripe was that by the last few episodes I was starting to tune out a bit. Other than that, this show gets a big recommendation.	8	Positive
I don't get this	I really don't know what all the fuss is about... The acting is strong (as expected from Korean actors and actresses) but most people are talking about the plot, which to me, is nothing special. Sure there are some plot twists but there are also way too many gaps for this to be called a "classic" or "masterpiece".	6	Positive
Lame	Acting 3/10. Story 1/10. Visuals 6/10. Feels like a B rated Hollywood movie from the 80s. Is it reddit or 4chan behind the hype again?	2	Negative
Overhyped and boring.	I don't know why people hype this up so much. Story is boring and predictable. The VIPs were bad actors, and the other characters were mostly annoying. As said in my first sentence I can't seem to understand why people like this so much and that it is rated an 8.3..	4	Negative

Table 4 is the result of adding a new feature, namely sentiment. Next, the classification process is carried out using the Support Vector Machine (SVM) method for the sentiment class.

Classification

The last process is classification with a comparison of 70% training data and 30% testing data to get the results of accuracy, precision, and recall. The classification process uses the SVM method with the selected kernel, which is linear. Support Vector Machine (SVM) is an important part of machine learning theory. SVM is very efficient for many applications in science and engineering, especially for classification (pattern recognition) problems (Ramadhan, N. G., & Khoirunnisa, A.2021). The SVM classification idea can be described as follows: suppose there are m samples of observations (training set), (*xi*, *yi*), *i = 1, 2, ..., m* where:

$$= (\quad 1, \ldots \quad) \in \tag{2}$$

Where is the d-dimensional feature of the sample *i* and *yi* ∈ *{-1, +1}* is the coded label class. If sample *xi* assigned to positive class, so *yi* is +1, and if assigned to negative class, so *yi* is -1. This training set can be separated by a hyperplane + = 0, where *w* is the weight vector and *b* is bias. The linear kernel formula can be seen in (3).

$$(\ , \ ) = \tag{3}$$

Linear kernels are very suitable for use on data that has many features such as text data. Kernel functions and parameters used in SVM analysis greatly affect the accuracy that will be generated (Ramadhan, N. G., & Khoirunnisa, A.2021).

RESULT

In this section, to see the results of accuracy, precision, and recall, calculations using confusion matrix (4), (5), and (6) are used.

$$= \frac{+}{+ \quad + \quad +} \tag{4}$$

$$/ \quad = (TP + FN) \tag{5}$$

*name of corresponding author

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

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, January 2022
DOI : https://doi.org/10.33395/sinkron.v7i1.11204
e-ISSN : 2541-2019
p-ISSN : 2541-044X

$$/ \quad = \left(\frac{TP}{TP + FP} \right) \tag{6}$$

Table 5 shows the results obtained using the proposed research method and other methods used.

Table 5
Result

Methods	Accuracy	Precision	Recall
Support Vector Machine	79%	75%	87%
Logistic Regression	67%	61%	82%

The results in table 5 show that the proposed method of SVM in the case of movie reviews using comments and star ratings on the website can produce higher accuracy, precision, and recall than other methods used. This happens because the SVM linear kernel is very suitable for use in data classification problems in the form of text. SVM is a method that is proven to be able to classify text data with high accuracy results as well. However, from the results of accuracy, precision, and recall, the precision value can still tend to be lower. This indicates that when repeated testing is carried out there are values whose results are very far from the truth.

CONCLUSION

Based on the results obtained in this study, it is proven that the purpose of analyzing sentiment on the comments of IMDB website users with the rating aspect can be done. The SVM classification model used in this study is also able to produce 79% accuracy, 75% precision, and 87% recall. SVM is very suitable for use in text data problems. For further research, additional aspects such as film genre can be added.

REFERENCES

Tripathi, A., & Trivedi, S. K. (2016, October). Sentiment analyis of Indian movie review with various feature selection techniques. In *2016 IEEE international conference on advances in computer applications (ICACA)* (pp. 181-185). IEEE.

Zamzami, F. N., & Adiwijaya, A. (2021). Analisis Sentimen Terhadap Review Film Menggunakan Metode Modified Balanced Random Forest dan Mutual Information. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *5*(2), 415-421.

Shivaprasad, T. K., & Shetty, J. (2017, March). Sentiment analysis of product reviews: a review. In *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 298-301). IEEE.

Zainuddin, N., & Selamat, A. (2014, September). Sentiment analysis using support vector machine. In *2014 international conference on computer, communications, and control technology (I4CT)* (pp. 333-337). IEEE.

Mahyarani, M., Adiwijaya, A., Al Faraby, S., & Dwifebri, M. (2021, July). Implementation of Sentiment Analysis Movie Review based on IMDB with Naive Bayes Using Information Gain on Feature Selection. In *2021 3rd International Conference on Electronics Representation and Algorithm (ICERA)* (pp. 99-103). IEEE.

Abidin, Z., Destian, W., & Umer, R. (2021, June). Combining support vector machine with radial basis function kernel and information gain for sentiment analysis of movie reviews. In *Journal of Physics: Conference Series* (Vol. 1918, No. 4, p. 042157). IOP Publishing.

Ding, Z., Qi, Y., & Lin, D. (2021, March). Albert-based sentiment analysis of movie review. In *2021 4th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)* (pp. 1243-1246). IEEE.

Dashtipour, K., Gogate, M., Adeel, A., Larijani, H., & Hussain, A. (2021). Sentiment analysis of persian movie reviews using deep learning. *Entropy*, *23*(5), 596.

Asriyanti, I. P., & Adiwijaya, A. I. (2018). On the feature selection and classification based on information gain for document sentiment analysis. *Applied Computational Intelligence and Soft Computing*, *2018*.

Ulfa, M. A., Irmawati, B., & Husodo, A. Y. (2018). Twitter Sentiment Analysis using Na¨ ive Bayes Classifier with Mutual Information Feature Selection. *Journal of Computer Science and Informatics Engineering (J-Cosine)*, *2*(2), 106-111.

Gandhi, U. D., Kumar, P. M., Babu, G. C., & Karthick, G. (2021). Sentiment Analysis on Twitter Data by Using Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM). *Wireless Personal Communications*, 1-10.

*name of corresponding author

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

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

Rahman, A., & Hossen, M. S. (2019, September). Sentiment analysis on movie review data using machine learning approach. In *2019 International Conference on Bangla Speech and Language Processing (ICBSLP)* (pp. 1-4). IEEE.

Mumtaz, D., & Ahuja, B. (2016, July). Sentiment analysis of movie review data using Senti-lexicon algorithm. In *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)* (pp. 592-597). IEEE.

Yessenov, K., & Misailovic, S. (2009). Sentiment analysis of movie review comments. *Methodology*, *17*, 1-7. Lee, S. H., Cui, J., & Kim, J. W. (2016). Sentiment analysis on movie review through building modified sentiment dictionary by movie genre. *Journal of intelligence and information systems*, *22*(2), 97-113.

IMDB Movie. https://www.imdb.com/title/tt10919420/reviews?ref_=tturv_ql_3. Accessed October 4, 2021.

Olken, F., & Rotem, D. (1986). Simple random sampling from relational databases.

Kannan, S., Gurusamy, V., Vijayarani, S., Ilamathi, J., Nithya, M., Kannan, S., & Gurusamy, V. (2014). Preprocessing techniques for text mining. *International Journal of Computer Science & Communication Networks*, *5*(1), 7-16.

Ramadhan, N. G., & Khoirunnisa, A. (2021). Klasifikasi Data Malaria Menggunakan Metode Support Vector Machine. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *5*(4), 1580-1584.

*name of corresponding author

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