Building of Informatics, Technology and Science (BITS)
Volume 4, No 1, Juni 2022 Page: 1−6
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i1.1468

Implementation Naïve Bayes Classification for Sentiment Analysis on Internet Movie Database

Samsir, Kusmanto, Abdul Hakim Dalimunthe, Rahmad Aditiya, Ronal Watrianthos*

Fakultas Teknik, Program Studi Teknik Informatika, Universitas Al Washliyah Labuhanbatu, Rantauprapat, Indonesia
Email: [1]samsirst111@gmail.com, [2]Kusnabara03@gmail.com, [3]abdulhakimdalimunthe@gmail.com, [4]Raditiya2428@gmail.com, [5,*]ronal.watrianthos@gmail.com
Email Penulis Korespondensi: ronal.watrianthos@gmail.com

Submitted: **03/04/2022**; Accepted: **22/06/2022**; Published**: 30/06/2022**

Abstract−A film review is a subjective opinion of someone who has different feelings about each film. As a result, film enthusiasts will struggle to assess whether the film meets their requirements. Based on these issues, sentiment analysis is the best way to fix them. Sentiment analysis, also known as opinion mining, is the study of assigning views or emotional labels to texts in order to determine if the text contains positive or negative thoughts. The Nave Bayes method was chosen because it can classify data based on the computation of each class's probability against objects in a given data sample. The scenario will compare the utilization of data that has been lemmatized with data that has not been lemmatized. The test compares data preparation stages with and without lemmatization. The best model was created utilizing data without lemmatization, 500 vector sizes, and Nave Bayes classification, with an accuracy of 78.96 percent and a f1-score of 78.81 percent. Changes in vector size affect the system's capacity to foresee positive and negative sentiments. The difference in accuracy and recall values shows that when vector size 300 is utilized, the precision and recall outcomes are lower than when vector size 500 is used.

Keywords: Naïve Bayes; Sentiment Analysis; Internet Movie Database

1. INTRODUCTION

A movie is a spectacle that may be enjoyed at leisure. There are numerous movies available today that may be viewed on the internet or at a theater. Movies that are streamed on the internet are often paid to watch so that future viewers may read comments from people who have watched the movie before watching it. A film review or film review is someone is a subjective opinion, and everyone will have a different perspective on each film. Each review of a film makes a distinct argument. Sentiment analysis can help to solve these issues. Sentiment analysis, also known as opinion mining, is the study of assigning views or emotional labels to texts in order to determine if the text reflects positive or negative opinions[1][2][3].

IMDb (Internet Movies Database) is a website that provides film reviews for films that have already been released. Today, IMDB is a popular website for viewing movie reviews. On this website, one may search for a film to watch by first reading the comments to select which film to watch based on the best or negative remarks. On the IMDB website, movie comments are many and varied; you may view comments depending on the star rating component. This makes it harder for users to analyze the remarks of other users[4][5]. Movie reviews help people decide if a film is worth watching and whether it is worth their time. A summary of all reviews for a film can aid consumers to make this decision by preventing them from wasting time reading all assessments. Critics frequently utilize film-rating websites and fee movies to assist visitors to decide if the picture is worth viewing. A movie review is a collection of facts that serves as an appraisal of each component of a film. The information in the film may be concluded as the audience's quality of a film experience, however, the rating of 'inappropriate' with the context sentence makes the film's rating 'not recommended.' This problem lends credence to the notion that a film should be classed based on its emotional content[6].

Many researchers have undertaken studies on sentiment analysis in nature using various classification methods and feature extraction. The sentiment analysis on Spider-Man: No Way movie review reveals 94 positive and 65 negative reviews. The pre-processing stage is critical for enhancing accuracy in the classification experiment. In the processed title dataset, Naive-Bayes is the best accurate classifier. In the original title and original comment datasets, SVM is the most accurate classifier[7]. The Long Short-Term Memory (LSTM) classifier is utilized in this work to analyze the sentiments of IMDb movie reviews. The Recurrent Neural Network (RNN) algorithm lies at the heart of it. To improve post-classification performance, the data is efficiently preprocessed and partitioned. The accuracy of classification performance is investigated. The best categorization accuracy was found to be 89.9 percent. It validates the feasibility of incorporating the provided method into contemporary text-based sentiment analyzers[8].

Other studies demonstrate that combining Machine Learning features (TF, TF-IDF) with Lexicon features (Positive-Negative word count, Connotation) yields better results in terms of accuracy and complexity when compared to classifiers such as SVM, Naive Bayes, KNN, and Maximum Entropy. The suggested approach clearly distinguishes between positive and negative reviews. Because knowing the context of the reviews is vital in classification, utilizing hybrid features aids in capturing the context of the movie reviews and so improves classification accuracy[9].

Other studies build sentiment analysis models on textual movie reviews using various feature extraction (count vectorization, TF-IDF, and Word2Vec) and feature selection (mutual information gain and Chi-square) techniques, and they compare the performance of various classification algorithms for building sentiment analysis models across several metrics. With TF-IDF Vectorization, Chi2 feature selection, and the SVM classification method, the result achieved the greatest accuracy of 90%. It was also shown that feature selection significantly decreases train test time for virtually all classification models while having no negative influence on other performance measures[10].

The goal of this research is to create a review sentiment analysis on IMDB films using the Nave Bayes classifier. The Nave Bayes technique was chosen because it can categorize data based on the computation of each class's probability against items in the provided data sample. Naive Bayes estimates a class's probability based on its qualities and selects the class with the highest probability. By assuming that each feature in the data is mutually exclusive, Nave Bayes classifies classes based on basic probabilities[11].

2. RESEARCH METHODOLOGY

2.1 Research Stages

Using the Word2Vec feature extraction method and the Naïve Bayes classification algorithm, this study constructs sentiment analysis on IMDB using film reviews. The flow of the system to be created is depicted in the diagram below.

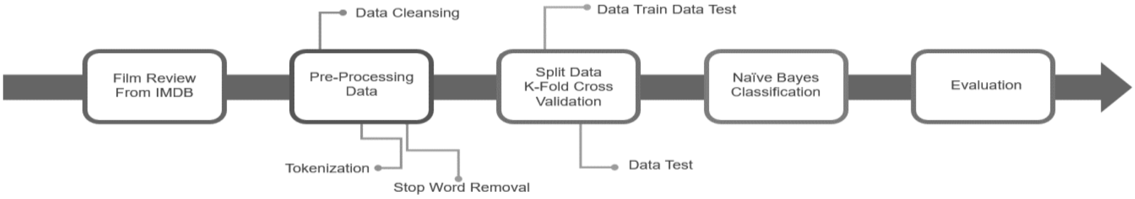

Figure 1. Research Stage

2.2 Data Set

The dataset for this study was derived from the IMDB Dataset of 50K Movie Reviews[12]. This is a binary sentiment classification dataset with significantly more data than prior benchmark datasets. We supply a set of 25,000 highly polar film reviews for training and another 25,000 for testing. So, using either classification or deep learning algorithms, forecast the quantity of favorable and negative reviews.

Table 1. Sample Data Set

Review	Sentiment
If you like original gut wrenching laughter you will like this movie. If you are young or old then you will love this movie, hell even my mom liked it.<br /><br />Great Camp!!!,	Positive
Besides being boring, the scenes were oppressive and dark. The movie tried to portray some kind of moral, but fell flat with its message. What were the redeeming qualities?? On top of that, I don't think it could make librarians look any more unglamorous than it did	Negative
It's terrific when a funny movie doesn't make smile you. What a pity!! This film is very boring and so long. It's simply painfull. The story is staggering without goal and no fun.<br /><br />You feel better when it's finished	Negative
It's this sort of movie that you try and imitate. By attempting to realise something... then flying through the air almost immediately. I'd like to do that and I know you would too!<br /><br />Great stuff!	Positive

2.2 Data Cleansing

At this point, any unnecessary symbols, numerals, html elements, or spaces are eliminated. The data will next go through the case folding procedure, where it will be changed to lowercase. Here's an example of a data purification method for film review data:

Table 2. Data Cleansing

Before	After
If you like original gut wrenching laughter you will like this movie. If you are young or old then you will love this movie, hell even my mom liked it.<br /><br />Great Camp!!!,	*If you like original gut wrenching laughter you will like this movie If you are young or old then you will love this movie hell even my mom liked it Great Camp*
Besides being boring, the scenes were oppressive and dark. The movie tried to portray some kind of moral, but fell flat with its message. What were the redeeming qualities?? On top of that, I don't think it could make librarians look any more unglamorous than it did	*Besides being boring, the scenes were oppressive and dark. The movie tried to portray some kind of moral, but fell flat with its message What were the redeeming qualities On top of that I don't think it*

Before	After
	could make librarians look any more unglamorous than it did
It's terrific when a funny movie doesn't make smile you. What a pity!! This film is very boring and so long. It's simply painfull. The story is staggering without goal and no fun.<br /><br />You feel better when it's finished	*It's terrific when a funny movie doesn't make smile you What a pity This film is very boring and so long. It's simply painfull The story is staggering without goal and no fun You feel better when it's finished*
It's this sort of movie that you try and imitate. By attempting to realise something... then flying through the air almost immediately. I'd like to do that and I know you would too!<br /><br />Great stuff!	*It's this sort of movie that you try and imitate By attempting to realise something then flying through the air almost immediately. I'd like to do that and I know you would too Great stuff*

2.3 Tokenization

Tokenization is the process of converting a word from a phrase into a token. Here's an example of a tokenization procedure for film review data:

Table 3. Example of Tokenization Process

Before	After
If you like original gut wrenching laughter you will like this movie. If you are young or old then you will love this movie, hell even my mom liked it.<br /><br />Great Camp!!!,	*"If"" you" "like" "original" "gut" "wrenching" "laughter" "you" "will" "like" "this" "movie" "If" "you" "are" "young" "or" "old" "then" "you" "will" "love" "this" "movie" "hell" "even" "my" "mom" "liked" "it" "Great" "Camp"*

2.4 Stopword Removal

The technique of identifying the form of a word dictionary given one of the inflectional alternatives, such as affixes or insertions, is known as lemmatization. In this work, we employed the NLTK library's WordNetLemmatizer to conduct lemmatization on English data. An example of the lemmatization procedure using film review data is as follows:

Table 4. Example of Stopword Removal

Before	After
If you like original gut wrenching laughter you will like this movie. If you are young or old then you will love this movie, hell even my mom liked it.<br /><br />Great Camp!!!,	*You like original gut wrenching laughter you will like this movie you young old then you will love this movie hell even my mom liked Great Camp*

2.5 Lemmatization

The technique of identifying the form of a word dictionary given one of the inflectional alternatives, such as affixes or insertions, is known as lemmatization. In this work, we employed the NLTK library's WordNetLemmatizer to conduct lemmatization on English data. An example of the lemmatization procedure using film review data is as follows:

Table 5. Example of Lemmatization

Before	After
Besides being boring, the scenes were oppressive and dark. The movie tried to portray some kind of moral, but fell flat with its message. What were the redeeming qualities?? On top of that, I don't think it could make librarians look any more unglamorous than it did	*Beside be boring, the scene were oppressive and dark. The movie tried to portray some kind of moral, but fell flat with it is message What were the redeeming qualities On top of that I don't think it could make librarian look any more unglamorous than it did*

2.6 Cross Validation

Cross validation is a technique for separating a dataset into training and test data. Cross validation will employ k-fold, which is divided into numerous portions, to split the dataset into training and test data. In this investigation, k-fold = 10 was used. If k is more than 10, the dataset will be partitioned into ten parts. Each dataset will include nine training data points and one test data point.

2.7 Word2Vec

Word2Vec is the current standard for natural language processing activities. A similar source of inspiration was discovered in current advanced deep neural networks' distributed embedding (word vectors). The improper mix of hyperparameters, on the other hand, might result in low-quality embeddings. The goal of this study is to empirically demonstrate that Word2Vec ideal hyper-parameter combinations exist and to analyze various combinations[13].

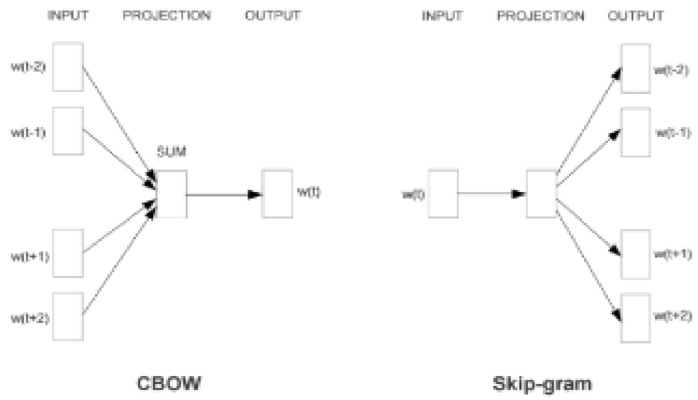

Figure 2. Model of Word2Ve

Word2Vec, created by Mikolov et al. (2013), is one of the continuous learning technologies used to generate word embedding. Word2Vec models are classified into two variants based on the amount of anticipated words: the Continuous Bag-of-Words (CBOW) model and the Skip-Gram model. Figure 2 depicts the Word2Vec model[14].

2.8 Naïve Bayes Clasifier

The naive Bayes classification method is based on Bayesian theory and uses statistical methods to predict the likelihood of membership in each class. Based on the Nave Bayes assumption, which states that the attribute value of one class will not impact the attribute value of another class (class-conditional independence)[2][15]. The total probability distribution is thus the product of the probability distributions of each data tuple, as illustrated in the equation below:

$$P(X|C_i) = \prod_{i=1}^{n} P(x_k|C_i)$$
$$= P(x_1|C_1) \times P(x_2|C_2) \times \ldots \times P(x_k|C_i) \qquad (1)$$

The created model will next be examined to establish the model's outcomes and performance level in categorizing film review data. A confusion matrix will be used to evaluate the model. Precision, recall, and f1-score will be calculated using the confusion matrix. The following formula is used to calculate precision, recall, and f1-score:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
$$precision = \frac{TP}{TP + FP}$$
$$recall = \frac{TP}{TP + FN}$$
$$f1 - score = 2 \times \frac{(precision \times recall)}{(precision + recall)} \qquad (2)$$

TP is true positive, TN is true negative, FP is false positive and FN is false negative[16].

3. RESULT AND DISCUSSION

The next step is to create a sentiment classification model with the Nave Bayes algorithm. The classification model developed will be assessed using various test scenarios to determine the ideal combination for producing the greatest results. The scenario will compare the utilization of data that has been lemmatized with data that has not been lemmatized. The test compares data preparation stages with and without lemmatization. The test makes use of a 500-point vector, a 5-point window, and an IMDb corpus dataset. The following table displays the test results for the first scenario.

Table 6. Test Result

Preprocessing	Accuracy	Precision	Recall	F1-Score
Lemmatization	78.54%	77.21%	78.81%	78.20%
Without Lemmatization	78.96%	77.99%	78.90%	78.81%

Table 6 demonstrates that the dataset without lemmatization performs better, with an accuracy of 78.54 percent, precision of 77.21 percent, recall of 78.81 percent, and f1-score of 78.20 percent. This is due to the lemmatization stage of the dataset removing words that do not match their meaning. Based on the precision results, it is clear that the system developed has a reasonable capacity to predict all positive sentiments from all true positive sentiments. While the recall results demonstrate that the algorithm can predict all negative sentiments from all real negative sentiments.

The next test is performed with data without a lemmatization procedure since it provides the greatest results. The efficacy of the Word2Vec size vector between 300, 400, and 500 with a window size of 5 was tested, and the results are displayed in table 7 below:

Table 7. Word2Vec Test Result

Vector Size	Accuracy	Precision	Recall	F1-Score
300	79.21%	79.56%	79.11%	79.10%
400	79.13%	79.69%	79.80%	79.12%
500	79.44%	79.65%	79.65%	79.77%

Table 7 demonstrates that Word2Vec with vector size 300 and window size 5 has the best performance with 79.21% accuracy, 79.56% precision, 79.11% recall, and 79.10% f1-score. The size of the vector affects the resultant performance since the higher the vector size, the more backpropagation happens in the hidden layer. However, a big vector size does not necessarily result in a good performance; this is due to the quantity of data utilized in developing the Word2Vec model. Changes in vector size have an impact on the system's ability to forecast positive and negative feelings. This is demonstrated by the difference in accuracy and recall numbers in table 7, where it can be seen that when vector size 300 is used, the precision and recall results are lower than when vector size 500 is used.

4. CONCLUSION

The following are the conclusions that may be derived from this study based on the findings of the investigation. The comparison of preprocessing scenarios with and without lemmatization yields varied outcomes. In this investigation, data without lemmatization performed better than data with lemmatization. Based on the results given above, the best English film review sentiment analysis system was created utilizing the corpus IMDb dataset without preprocessing lemmatization, with a vector size of 500 and a window size of 5 producing rather decent results. In order to determine how Naive Bayes compares to other algorithms, further research should be needed.

REFERENCES

[1] H. S. Batubara, Ambiyar, Syahril, Fadhilah, and R. Watrianthos, "Sentiment Analysis of Face-To-Face Learning Based on Social Media," *J. Pendidik. Teknol. Kejuru.*, vol. 4, no. 3, 2021.

[2] Samsir, Ambiyar, U. Verawardina, F. Edi, and R. Watrianthos, "Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes," *J. Media Inform. Budidarma*, vol. 5, no. 1, pp. 157–163, 2021, doi: 10.30865/mib.v5i1.2604.

[3] H. S. Batubara, M. Giatman, W. Simatupang, and R. Watrianthos, "Pemetaan Bibliometrik Terhadap Riset pada Sekolah Menengah Kejuruan Menggunakan VOSviewer," *Edukatif J. Ilmu Pendidik.*, vol. 4, no. 1, pp. 233–239, 2022.

[4] N. G. Ramadhan and T. I. Ramadhan, "Analysis Sentiment Based on IMDB Aspects from Movie Reviews using SVM," *Sinkron*, vol. 7, no. 1, pp. 39–45, Jan. 2022, doi: 10.33395/sinkron.v7i1.11204.

[5] 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. 1733–1741, 2017.

[6] M. Mahyarani, A. Adiwijaya, S. Al Faraby, and M. Dwifebri, "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)*, Jul. 2021, pp. 99–103. doi: 10.1109/ICERA53111.2021.9538763.

[7] P. H. Gunawan, T. D. Alhafidh, and B. A. Wahyudi, "The Sentiment Analysis of Spider-Man: No Way Home Film Based on IMDb Reviews," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 6, no. 1, pp. 177–182, Feb. 2022, doi: 10.29207/resti.v6i1.3851.

[8] S. M. Qaisar, "Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory," in *2020 2nd International Conference on Computer and Information Sciences (ICCIS)*, Oct. 2020, pp. 1–4. doi: 10.1109/ICCIS49240.2020.9257657.

[9] K. Kumar, B. S. Harish, and H. K. Darshan, "Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method," *Int. J. Interact. Multimed. Artif. Intell.*, vol. 5, no. 5, p. 109, 2019, doi: 10.9781/ijimai.2018.12.005.

[10] G. Karak, S. Mishra, A. Bandyopadhyay, P. R. S. Rohith, and H. Rathore, "Sentiment Analysis of IMDb Movie Reviews: A Comparative Analysis of Feature Selection and Feature Extraction Techniques," in *Hybrid Intelligent Systems*, 2022, pp. 283–294. doi: 10.1007/978-3-030-96305-7_27.

[11] R. Novendri, A. S. Callista, D. N. Pratama, and C. E. Puspita, "Sentiment Analysis of YouTube Movie Trailer Comments Using Naïve Bayes," *Bull. Comput. Sci. Electr. Eng.*, vol. 1, no. 1, pp. 26–32, Jun. 2020, doi: 10.25008/bcsee.v1i1.5.

[12] L. N, "IMDB Dataset of 50K Movie Reviews," *kaggle.com*, 2019. https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews (accessed Apr. 02, 2022).

[13] T. Adewumi, F. Liwicki, and M. Liwicki, "Word2Vec: Optimal hyperparameters and their impact on natural language processing downstream tasks," *Open Comput. Sci.*, vol. 12, no. 1, pp. 134–141, Mar. 2022, doi: 10.1515/comp-2022-0236.

[14] Alvi Rahmy Royyan and Erwin Budi Setiawan, “Feature Expansion Word2Vec for Sentiment Analysis of Public Policy in Twitter,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 6, no. 1, pp. 78–84, Feb. 2022, doi: 10.29207/resti.v6i1.3525.

[15] Samsir *et al.*, “Naives Bayes Algorithm for Twitter Sentiment Analysis,” *J. Phys. Conf. Ser.*, vol. 1933, no. 1, p. 012019, Jun. 2021, doi: 10.1088/1742-6596/1933/1/012019.

[16] R. Watrianthos, S. Suryadi, D. Irmayani, M. Nasution, and E. F. S. Simanjorang, “Sentiment analysis of traveloka app using naïve bayes classifier method,” *Int. J. Sci. Technol. Res.*, vol. 8, no. 7, pp. 786–788, 2019, doi: 10.31227/osf.io/2dbe4.