 ISSN 2579-7298

Improving Performance Sentiment Analysis Movie Review Film using Random Forest with Feature Selection Information Gain

Vinsent Brilian Adiguna[1], Muslihul Aqqad[2], Purwanto[3*], Jaluanto Sunu PT[4], Honorata Ratnawati DP[5]

[1,2,3] Department of Faculty Informatics Engineering, University Dian Nuswantoro, Semarang, Indonesia
[4,5] Department of Faculty Economics and Business, University 17 August 1945 Semarang, Semarang, Indonesia
[1]p31202202517@mhs.dinus.ac.id, [2]p31202202553@mhs.dinus.ac.id, [3]purwanto@dsn.dinus.ac.id, [4]jaluanto@untagsmg.ac.id,[5]honorata-ratnawati@untagsmg.ac.id

ARTICLE INFO

Article history
Received
Revised
Accepted

Keywords
Random Forest,
Information Gain,
Feature Selection,
Sentiment Analysis.

ABSTRACT

Sentiment analysis in film reviews is an important task to understand the audience's opinion towards a cinematic work. However, the complexity and subjectivity of language in film reviews pose a challenge. This research explores the application of Random Forest algorithm, an ensemble learning method, to perform sentiment classification on film reviews. Random Forest is built from a set of decision trees, each of which provides a prediction, and the final result is obtained from majority voting. This approach has the advantage of handling overfitting data. This research uses 500 review datasets along with positive and negative sentiment labels. The review text is represented as Information Gain and TF-IDF features to model the weight of each word. The Random Forest model is then trained using these features to predict sentiment labels. The performance of the model is evaluated using metrics such as accuracy, precision, recall and f1-score. The experimental results show that Random Forest is able to achieve 95.20% accuracy in sentiment classification of film reviews, surpassing the Support Vector Machine classification algorithm which in previous studies only achieved 92%. These findings provide a new perspective on the benefits of ensemble learning in sentiment analysis and its potential application in other domains such as marketing and public opinion analysis.

1. Introduction

Sentiment analysis of movie reviews has attracted a lot of interest in recent years due to the increasing number of movie reviews available online [1]. These reviews provide valuable insights into audience opinions and preferences towards a cinematic work. However, manually analyzing sentiment from review texts can be a challenging and time-consuming task, especially with large data volumes [2]. Therefore, automated sentiment analysis techniques based on machine learning become essential to help understand audience opinions efficiently and effectively.

As the volume of online movie reviews increases, there is a need to efficiently and accurately analyze and understand the sentiment contained in these reviews. Sentiment analysis becomes an essential tool in this context. Sentiment analysis is a computational process for identifying and categorizing opinions expressed in text, primarily to determine whether the author's attitude toward a topic, product, or in this case a movie, is positive, negative, or neutral.

In the context of movie reviews, sentiment analysis has several important benefits. First, for the film industry, sentiment analysis can provide valuable insights into audience reception of certain films, helping film producers and distributors make decisions regarding marketing and distribution strategies. Second, for streaming platforms and movie review websites, sentiment analysis can help in recommending movies to users based on their preferences. Third, for viewers, sentiment analysis results can be a quick guide in choosing movies that suit their tastes.

However, analyzing sentiment in movie reviews is not an easy task. The complexity of language, the use of irony and sarcasm, and the variation in writing styles between reviewers make this task a challenge in the field of natural language processing (NLP). Furthermore, the large volume of movie review data generated every day requires an approach that is not only accurate but also computationally efficient.

To overcome these challenges, various machine-learning methods have been applied to the task of movie review sentiment classification. One method that shows great potential is Random Forest. Random Forest is an ensemble learning algorithm that combines multiple decision trees to produce more accurate and stable predictions.

Several previous studies have discussed the *Naïve Bayes* algorithm that is better at classifying movie review sentiment analysis using weighting techniques TF-IDF feature and selection features using chi-square. This study focuses on the problem of improving accuracy by adding chi-square selection features and feature weighting using Term Frequency Inverse Document Frequency (TF-IDF). The results show that chi-square and TF-IDF improve the accuracy of the *Naïve Bayes* classifier to 83% better [3].

Jagdale and Deshmukh's study [4] used several datasets to train and test the SVM algorithm, which produced an accuracy of 89.98%. This result shows that SVM is a good choice for sentiment classification tasks. However, this study shows that the accuracy can be improved by considering more sentence forms. The results of this study have potential applications in improving product sentiment analysis, and future studies using more sophisticated extraction features, overall, this study understands and improves sentiment analysis in product reviews.

One of the machine learning algorithms that has proven to be effective in various text classification tasks is Random Forest [5]. Random Forest is an ensemble learning method that combines several decision trees to form a more robust and accurate model. Each decision tree in the ensemble is built using a random subset of the features and training data, thereby reducing the risk of overfitting and increasing model generalization [6].

In Sel et al.'s research [7], previous research explored the use of the Random Forest algorithm using the Information Gain criterion to select the most discriminative features in movie review sentiment classification. Information Gain is a metric that measures how much information is obtained to make a classification after observing a particular feature value. By selecting features that have the highest information gain, we can focus the model on the most relevant aspects in distinguishing positive and negative sentiment in movie reviews.

In the field of NLP, one of the important stages is text pre-processing. Text pre-processing involves a series of techniques to clean and prepare text data so that it can be processed more effectively by algorithms. Some commonly used text pre-processing techniques include case normalization, tokenization, removing punctuation, removing stop words, stemming, removing numbers, removing extra whitespace, text cleaning, and normalization.

This study aims to evaluate the performance of the Random Forest algorithm with information gain in sentiment analysis of movie reviews. This study uses 500 datasets, 250 positive sentiment datasets, and 250 negative sentiment datasets. This dataset has also been used in previous research by Sutriawan et al [8], as well as comparing the performance of the model with other classification algorithms such as *Naïve Bayes* and *Support Vector Machine*. The results of this study are expected to contribute to the deployment of a more accurate and efficient sentiment analysis system, especially in the domain of movie reviews.

Table 1. Literature Review

Author's	Feature Selection or Extraction	Classifier	Best Accuracy	Years

[3]	Chi-Square, TF-IDF	Naive Bayes	83%	2022
[6]	Bag of Words	Naive Bayes	71%	2023
[6]	Bag of Words	SVM	73%	2023
[9]	Ensemble Combined (Information Gain, Chi Square, Gini Index)	Random Forest	87,4%	2018
[10]	TF-IDF	SVM + Best Parameter	91,63%	2022
[8]	TF-IDF, Ngram, Information Gain	Linier SVM + Unigram + Information Gain	92%	2023

2. Methods

The proposed research method is anadaptation of previous research [8] which focuses on *Pre-Processing* followed by *feature extraction*, division of training data and test data, so that it can be classified with the *Naïve Bayes*, *Support Vector Machine* (SVM), *Random Forest* algorithm. Furthermore, a *comparison evaluation matrix* will be carried out.

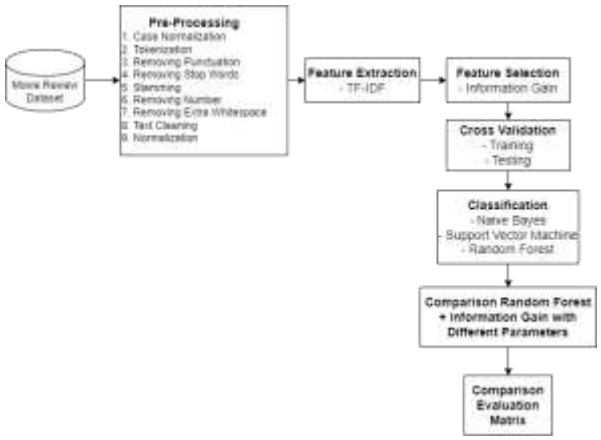

Fig. 1. Proposed method

Based on Figure 1, pre-processing techniques are very important starting from case normalization (changing uppercase letters to lowercase), tokenization (separating text into arrays), removing punctuation, removing stop words (removing common words such as; which, this), stemming (changing text to its basic form), removing number (removing numbers in the text), removing extra whitespace (removing too many spaces), text cleaning (removing characters such as symbols), normalization (normalizing text slang or abbreviations to standard form), after the text is cleaned, feature extraction is carried out using the TF-IDF method. The training data used is taken from a collection of dataset reviews that have been labeled positive and negative. As much as 90% of the entire dataset is used to create a sentiment analysis model and 10% for test data.

2.1 Data Collection

The initial stages in the process of analyzing film review sentiment. The quality and characteristics of the collected dataset will greatly affect the performance of the classification model and the validity of the research results [11]. The film review data used in this study were obtained from previous studies by Nurdiansyah et al., and Sutriawan et al. [8]. This dataset is in the form of Indonesian language film review data. The amount of data used is 500 consisting of 250 positive and 250 negative sentiments. This dataset is used to test the performance of the Random Forest classification algorithm by dividing it into two types, namely data testing and data training. The data training used is taken from a collection of dataset reviews that have been labeled positive and negative as much as 90% of the total dataset is used to form a sentiment analysis model and 10% for test data.

2.2 Data Pre-Processing

In NLP processing, the pre-processing stage is crucial for cleaning and preparing text before further processing [12]. Some techniques for cleaning and preparing text data so that it can be processed more effectively include case normalization (converting all letters in the text to lowercase) to standardize the text format [12]. Tokenization is done to divide text into tokens, such as words,

numbers, or punctuation [13]. Removing Punctuation is the process of removing all punctuation from the text while removing stop words is the process of removing common words that are less meaningful in a particular context such as 'this', 'and', 'that', 'or'. Stemming is the process of converting words that come from the same root word into a base word form by removing affixes such as prefixes and suffixes [13]. Removing Numbers is the process of removing all numbers from the text while removing extra whitespace is the process of removing extra spaces that are not needed in the text. Text cleaning is the process of cleaning text from noise such as 'URLs', 'emails', and 'emoticons'. Normalization is the process of transforming text into a more consistent and standardized form, such as removing text in brackets and converting acronyms to a more explicit form [14].

2.3 Feature Extraction

After the text is pre-processed, feature extraction is performed using a popular word weighting method in text processing, namely TF-IDF which calculates the weight of each word based on the frequency of the word appearing in the document and corpus [12]. TF-IDF is a feature extraction method commonly used in natural language processing. This method combines two important concepts, Term Frequency (TF) to measure how often a word appears in a document. The more often the word appears, the higher its value and Inverse Document Frequency (IDF) is to measure the importance of a word in the entire corpus, and words that rarely appear are considered more important. TF-IDF gives higher weight to words that appear frequently in a particular document but rarely appear in other documents in the corpus [15]. In the context of movie review sentiment analysis, TF-IDF can help identify keywords that are more relevant to positive or negative sentiment, improving classification accuracy [16].

2.4 Feature Selection

Feature selection performs a technique to select features or features in a data that have no relevance or are redundant in a set of data. Feature selection using *Information Gain* is one method of feature selection, in the *Information Gain* process the features will be ranked, the largest feature ranking is the most relevant feature and has a strong connection with the related data set [17]. This technique ranks features by calculating the *Entropy* of one class before and after conducting the observation process of the features in the same data[18].

2.5 Cross Validation

Model evaluation techniques that are used to assess how the results of a statistical analysis will generalize to an independent dataset. This technique is mainly used in contexts where the goal is prediction and one wants to estimate how accurate the model has been created [19]. In the study where the *K-Fold Cross Validation* method was used the dataset is divided into K-subsets of equal size. The model is trained on k-1 subsets and tested on the remaining subsets. This process is repeated as many times as K inputs are entered, with each subset used once as test data [20]. The dataset is divided into training data and test data using the cross-validation technique, a model evaluation method that involves partitioning the data into the most exclusive subsets, where the model is trained on a particular subset and evaluated on another subset [21], using K-Fold Validation K=10.

2.6 Clasification

One of the machine learning tasks that predicts categorical labels for a given input. In the context of movie review sentiment analysis the goal is to categorize reviews into positive or negative sentiment [22]. The *Naïve Bayes* algorithm is a probabilistic classification algorithm based on *Bayes' Theorem* with the assumption of independence between features; although this assumption is ‘naïve’ this algorithm often works very well in practice, especially for text classification. [23]. In sentiment analysis, *naïve bayes* calculates the probability of each word in a review to determine the overall sentiment [24]. Algoritma *SVM* sangat efektif untuk klasifikasi teks dan analisis sentimen, dengan fleksibelitas penggunaan fungsi kernel dan mengurangi overfitting [25]. SVM can also handle language complexity well and often produce high accuracy [25]. The Random Forest algorithm is an ensemble method that uses many decision trees to perform classification. Each tree provides a prediction, and the class with the most votes becomes the model output [26]. *Random Forest* can handle many features without overfitting and can capture complex patterns in text and often performs well [27]. In this stage, the movie review data has been pre-processed and will be

classified based on its positive and negative sentiments. This study proposes using the TFIDF and Information Gain feature extraction models to produce the most relevant features for the movie review data. In the classification stage, three machine learning algorithms are used, namely *Naïve Bayes*, *Support Vector Machine*, and *Random Forest*.

3. Results and Discussions

In this section, we will present the experimental results and analysis of three classification algorithms used for movie review sentiment analysis. The results obtained reflect the performance of each algorithm in classifying movie review sentiment into positive and negative categories. The experiments were conducted using a movie review dataset that had gone through the pre-processing and feature extraction stages using *TF-IDF*. The performance of each algorithm is evaluated using the K-Fold Cross Validation method to ensure the results obtained. The Evaluation Matrix used is accuracy which provides an overview of the classification capabilities of each algorithm. In the results section, data visualization is presented in the form of a bar chart to test the significance of differences in performance between algorithms. The limitations of this study is the field of film review sentiment analysis and comparison of classification algorithms.

This section presents the results of the matrix evaluation for the classification model, and compares different partition ratios with the Naïve Bayes, SVM, Random Forest algorithms presented in the form of tables and graphics.

3.1. Naïve Bayes' Model Test Result

The data was obtained from previous research by Nurdiansyah and Sutriawan et al. [8] by entering data with positive and negative labels, and then tokenization was carried out to divide the text into tokens, such as words, numbers, or punctuation. *The Cross Validation* section consists of training and testing. The training section includes the application of the *Naïve Bayes* algorithm while the testing section includes the application of the model as a performance operator to calculate the level of accuracy of the applied algorithm.

The Cross Validation value given will be the focus of *the Naïve Bayes* test. This study was conducted 10 times with validation values ranging from 1–10x. This test randomly divides the data into 10 parts. In this part of the test, it is known that the accuracy level is 69%. Details of the resulting opinion classification can be found in Table 2 of the *Naïve Bayes* algorithm confusion matrix table.

Table 2. The Result of *Confusion Matrix Naïve Bayes* Algorithm

	true Positive	true Negative
pred Positive	178	83
pred Negative	72	167

Table 2 shows 178 reviews that are considered positive and 167 reviews that are considered negative. In addition, there are errors in classifying 72 reviews that should be considered positive and 83 reviews that should be considered negative. Looking at the prediction classification results the accuracy level of the Naïve Bayes model is 69% with K-Fold Cross Validation=10.

3.2. 3.2 SVM Model Test Result

To optimize the accuracy level of the Naïve Bayes model, the classic SVM algorithm was tested on the kernel section using the Dot kernel, and Cross Validation K-Fold=10. The results of the classic SVM test can be seen in Table 3.

Table 3. The Result of *Confusion Matrix SVM* Algorithm

	true Positive	true Negative
pred Positive	234	29
pred Negative	16	221

Table 3 shows 234 reviews that are considered positive and 221 reviews that are considered negative. In addition, there are errors in classifying 16 reviews that should be considered positive and 29 reviews that should be considered negative. Looking at the prediction classification results the accuracy level of the *Support Vector Machine* model is 91% with K-Fold Cross Validation=10.

3.3. 3.3 Random Fores's Model Test Result

The results of this study show an evaluation matrix of the comparison of criteria from the random forest model with different criteria of gain ratio, information gain, gini index with a number of trees of 50 trees, and a max depth of 10.

Table 4. The Result of *Confusion Matrix* Random Forest Gain Ratio

	true Positive	true Negative
pred Positive	238	16
pred Negative	12	234

The test results are shown in Table 4. There are 238 reviews that are considered positive and 234 reviews that are considered negative. In addition, there is an error in determining the number of 12 reviews that should be considered positive and 16 reviews that should be considered negative. Looking at the classification results, it found an increase in accuracy to 94.40% with the *Random Forest* model with *gain ratio* parameters and *cross validation K-Fold*=10.

Table 5. The Result of *Confusion Matrix* Random Forest Information Gain

	true Positive	true Negative
pred Positive	237	11
pred Negative	13	239

The test results are shown in Table 5. 237 reviews were considered positive and 239 reviews were considered negative. In addition, there was an error in determining the number of review classifications that should be considered positive reviews totaling 13 reviews and 11 reviews that should be considered negative. Looking at the results of the classification that has been carried out, there is an increase in accuracy to 95.20% with the *Random Forest* model with *information gain* parameters and *cross validation K-Fold*=10.

Table 6. The Result of *Confusion Matrix* Random Forest Gini Index

	true Positive	true Negative
pred Positive	240	15
pred Negative	10	235

The test results are shown in Table 6. 240 reviews are considered positive and 235 reviews are considered negative. In addition, there is an error in determining the number of review classifications that should be considered positive reviews totaling 10 reviews and 15 reviews that should be considered negative. Looking at the results of the classification that has been done, there is an accuracy of 95.00% with the *Random Forest* model with the *gini index* parameter and *cross validation K-Fold*=10.

Comparison of criteria using the random forest algorithm produces results as shown in Figure 1 below.

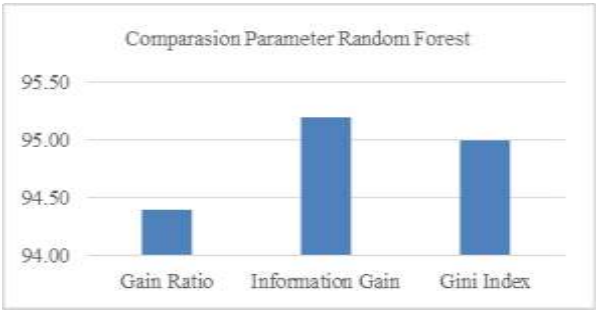

Fig. 2. Comparasion Parameter Random Forest

In Figure 1. The gain ratio parameter has an accuracy of 94.40% compared to the information gain parameter which gets an accuracy of 95.20% greater than the accuracy of the Gini index parameter which only produces 95.00%. This shows that the Random Forest algorithm with a number of trees of 50 trees, and a depth of 10, and the information gain parameter produces more accurate accuracy than the gain ratio parameter and the Gini index.

Based on the sentiment analysis testing stage using the Naïve Bayes, SVM, and Random Forest algorithms, several different accuracy levels were obtained. The Random Forest algorithm with the information gain parameter has the best performance with an accuracy of 95.20%, outperforming the Naïve Bayes and SVM algorithms. The SVM algorithm is in second place with an accuracy of 91%, this shows quite good performance in classifying sentiment. The Naïve Bayes algorithm has an accuracy of 69% which is relatively low compared to the other two algorithms.

Fig. 3. Comparasion of 3 Algorithm

Figure 2 shows that Random Forest has the highest accuracy of 95.20% compared to the accuracy of Naïve Bayes 69% and SVM 91% for the case of movie reviews with a dataset of 250 positive sentiments and 250 negative sentiments.

It didn't stop there, the researchers carried out further research by conducting research experiments by comparing the random forest algorithm with several parameters such as max depth and number of trees and different criteria such as the following results.

Table 7. Random Forest Results + Information Gain with a Maximum Depth of 5 and Different Parameters

Number of Tree	Maximal Depth	Information Gain	Gini Index	Gain Ratio
10	5	78.20%	74.80%	75.60%
20	5	85.60%	82.40%	78.40%
30	5	86.80%	86.00%	84.20%
40	5	88.60%	87.40%	85.80%
50	5	90.40%	89.40%	87.80%

The results from Table 7 show that with different numbers of trees (10,20,30,40,50) and with the same max a depth of 5 and cross validation with K=10 shows that Random Forest with information gain feature selection added criteria Information gain is still very superior compared to the Gini

index and gain ratio criteria with the highest accuracy results with a number of trees of 50 trees and a maximum depth of 5 producing an accuracy of 90.40%.

Researchers also experimented with a maximum depth of 10 and with different numbers of trees, the results are as in table 8 below.

Table 8. Hasil Random Forest + Information Gain with Maximal Depth 10 and Different Parameters

Number of Tree	Maximal Depth	Information Gain	Gini Index	Gain Ratio
10	10	83.80%	86%	82.40%
20	10	91%	89.80%	88.20%
30	10	92.60%	92.80%	90.40%
40	10	94%	94.80%	92.60%
50	10	95.20%	95.%	94.40%

The results from Table 8 show that with a number of trees of 10 and a maximum depth of 10 and a criterion Gini index, it succeeded in getting an accuracy of 86%. This shows that it can beat the criterion information gain which got an accuracy of 83.80% and the gain ratio which only got an accuracy of 82. 40%. The number of trees of 20 with a maximum depth of 10 shows that the information gain results have greater accuracy than the Gini index and gain ratio, which is 91%. The number of trees is 30 with a maximum depth of 10 and the criterion Gini index gets an accuracy result of 92.80% which is slightly different from the criterion information gain which gets a result of 92.60% and the gain ratio which only gets 90.40%, and the number of trees is 40 and the maximum depth 10 highest accuracy got a result of 94.80% with the Gini index criterion. The final test was carried out using a number of trees of 50 and a maximum depth of 10, showing that the criterion information gain accuracy was 95.20%, the Gini index criterion had an accuracy of 95%, as did the criterion gain ratio which only got a result of 94.40%. These tests were all carried out using cross validation with K=10

Based on the results of the research above, a comparison was carried out with previous researchers as in the following results in table 9. An important finding is that the combination of TFIDF, Selection Feature Information Gain, and Random Forest with the Information Gain criteria resulted in a significant increase in the classification of performance film reviews. with high accuracy.

Table 9. Performance comparasion of algorithms based on accuracy

Author's	Feature Selection or Extraction	Classifier	Best Accuracy
[3]	Chi-Square, TF-IDF	Naive Bayes	83%
[6]	Bag of Words	Naive Bayes	71%
[6]	Bag of Words	SVM	73%
[9]	Ensemble Combined (Information Gain, Chi Square, Gini Index)	Random Forest	87,4%
[10]	TF-IDF	SVM + Best Parameter	91,63%
[8]	TF-IDF, Ngram, Information Gain	Linier SVM + Unigram + Information Gain	92%
Purpose Method	TF-IDF, Information Gain	Random Forest	95,20%

4. Conclusions

Based on these results, Random Forest with feature selection Information Gain and criterion information gain is recommended as the best algorithm in the case of movie review sentiment analysis. This is because Random Forest has a better ability to capture complex patterns and relationships in text data, resulting in more accurate sentiment classification, but keep in mind that the selection of the best algorithm can vary depending on the type of data and the amount of data.

References

[1] B. Liu, "Sentiment analysis: Mining opinions, sentiments, and emotions," 2020, *Cambridge university press*. [Online]. Available: http://tcci.ccf.org.cn/conference/2014/ppts/adl/adl52-l3.pdf

[2] K. Ravi and V. Ravi, "A novel automatic satire and irony detection using ensembled feature selection and data mining," *Knowledge-Based Syst.*, vol. 120, pp. 15–33, 2017, doi: 10.1016/j.knosys.2016.12.018.

[3] A. Falasari and M. A. Muslim, "Optimize Naïve Bayes Classifier Using Chi Square and Term Frequency Inverse Document Frequency For Amazon Review Sentiment Analysis," *J. Soft Comput. Explor.*, vol. 3, no. 1, pp. 31–36, 2022, doi: 10.52465/joscex.v3i1.68.

[4] R. S. Jagdale, V. S. Shirsat, and S. N. Deshmukh, "Sentiment analysis on product reviews using machine learning techniques," *Cogn. Informatics Soft …*, 2019, doi: 10.1007/978-981-13-0617-4_61.

[5] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "VSURF: An R Package for Variable Selection Using Random Forests," *R J.*, vol. 7, no. 2, p. 19, 2015, doi: 10.32614/rj-2015-018.

[6] S. Mehla, "Sentiment Analysis of Movie Reviews using Machine Learning Classifiers," *Int. J. Comput. Appl.*, vol. 182, no. 50, pp. 25–28, Apr. 2019, doi: 10.5120/ijca2019918756.

[7] İ. Sel, A. Karci, and D. Hanbay, "Feature Selection for Text Classification Using Mutual Information," in *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, IEEE, Sep. 2019. doi: 10.1109/idap.2019.8875927.

[8] Sutriawan, Muljono, Khairunnisa, Z. Alamin, T. A. Lorosae, and S. Ramadhan, "Improving Performance Sentiment Movie Review Classification Using Hybrid Feature TFIDF, N-Gram, Information Gain and Support Vector Machine," *Math. Model. Eng. Probl.*, vol. 11, no. 2, pp. 375–384, 2024, doi: 10.18280/mmep.110209.

[9] M. Ghosh and G. Sanyal, "An ensemble approach to stabilize the features for multi-domain sentiment analysis using supervised machine learning," 2018, *Springer*. doi: 10.1186/s40537-018-0152-5.

[10] H. Mustakim and S. Priyanta, "Aspect-Based Sentiment Analysis of KAI Access Reviews Using NBC and SVM," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 16, no. 2, p. 113, 2022, doi: 10.22146/ijccs.68903.

[11] N. Bansal and H. Choudhary, "Charting the Trajectory of Digital Literacy Research: A Review of Research Topics, Publication Venues, and Top Cited Papers," *Int. J. Learn. Technol.*, vol. 1, no. 1, 2023, doi: 10.1504/ijlt.2023.10060691.

[12] W. Wagner, "Steven Bird, Ewan Klein and Edward Loper: Natural Language Processing with Python, Analyzing Text with the Natural Language Toolkit: O'Reilly Media, Beijing, 2009, ISBN 978-0-596-51649-9," *Lang. Resour. Eval.*, vol. 44, no. 4, pp. 421–424, 2010, doi: 10.1007/s10579-010-9124-x.

[13] P. Willett, "The Porter stemming algorithm: then and now," *Program*, vol. 40, no. 3, pp. 219–223, 2006, doi: 10.1108/00330330610681295.

[14] J. Han, M. Kamber, and J. Pei, "Data Preprocessing," in *Data Mining*, Elsevier, 2012, pp. 83–124. doi: 10.1016/b978-0-12-381479-1.00003-4.

[15] G. Salton, A. Singhal, M. Mitra, and C. Buckley, "Automatic text structuring and summarization," *Inf. Process. & Manag.*, vol. 33, no. 2, pp. 193–207, 1997, doi: 10.1016/s0306-4573(96)00062-3.

[16] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02*, in EMNLP '02. Association for Computational Linguistics, 2002. doi: 10.3115/1118693.1118704.

[17] S. Chormunge and S. Jena, "Efficient Feature Subset Selection Algorithm for High Dimensional Data," *Int. J. Electr. Comput. Eng.*, vol. 6, no. 4, p. 1880, Aug. 2016, doi: 10.11591/ijece.v6i4.9800.

[18] D. P. Utomo and M. Mesran, "Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung," *J. MEDIA Inform. BUDIDARMA*, vol. 4, no. 2, p. 437, Apr. 2020, doi: 10.30865/mib.v4i2.2080.

[19] J. S. Cavenaugh, "Bootstrap Cross-validation Improves Model Selection in Pharmacometrics," *Stat. Biopharm. Res.*, vol. 14, no. 2, pp. 168–203, Nov. 2020, doi: 10.1080/19466315.2020.1828159.

[20] M. Stone, "Cross-Validatory Choice and Assessment of Statistical Predictions (With Discussion)," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 38, no. 1, pp. 102–102, Sep. 1976, doi: 10.1111/j.2517-6161.1976.tb01573.x.

[21] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-Validation," in *Encyclopedia of Database Systems*, Springer New York, 2018, pp. 677–684. doi: 10.1007/978-1-4614-8265-9_565.

[22] D. K. Y. Chiu, "BOOK REVIEW: 'PATTERN CLASSIFICATION', R. O. DUDA, P. E. HART and D. G. STORK, Second Edition," *Int. J. Comput. Intell. Appl.*, vol. 01, no. 03, pp. 335–339, Sep.

2001, doi: 10.1142/s1469026801000251.

[23] C. D. Manning, "Natural Language Processing, Statistical Approaches to," 2006, *Wiley*. doi: 10.1002/0470018860.s00080.

[24] B. Liu, "Sentiment Analysis and Opinion Mining," *Synth. Lect. Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, 2012, doi: 10.2200/s00416ed1v01y201204hlt016.

[25] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features," in *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 1998, pp. 137–142. doi: 10.1007/bfb0026683.

[26] R. Wetteland, K. Engan, T. Eftestøl, V. Kvikstad, and E. Janssen, "Multiclass Tissue Classification of Whole-Slide Histological Images using Convolutional Neural Networks," in *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods*, SCITEPRESS - Science and Technology Publications, 2019. doi: 10.5220/0007253603200327.

[27] G. Tsoumakas and I. Katakis, "Multi-Label Classification: An Overview," *Int. J. Data Warehous. Min.*, vol. 3, no. 3, pp. 1–13, 2007, doi: 10.4018/jdwm.2007070101.