International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Improving Performance Sentiment Movie Review Classification Using Hybrid Feature TFIDF, N-Gram, Information Gain and Support Vector Machine

Check for updates

Sutriawan[1,2], Muljono[1]*, Khairunnisa[2], Zumhur Alamin[2], Teguh Ansyor Lorosae[2], Sahrul Ramadhan[2]

[1] Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang 50131, Indonesia
[2] Department of Computer Science, Universitas Muhammadiyah Bima, Bima 84116, Indonesia

Corresponding Author Email: muljono@dsn.dinus.ac.id

ABSTRACT

The use of online movie streaming media has increased significantly, particularly among movie enthusiasts. However, fan comments are frequently informal and comprise informal language, subjectivity, and contexts that reflect their preferences. A significant challenge in sentiment analysis of movie reviews is how to classify sentiments in reviews that are often unstructured and subjective. This study aims to improve the accuracy of sentiment classification in movie reviews by proposing several methods, including a hybrid TF-IDF+N-Gram model that can extract pertinent information from word and phrase sequences in reviews. Then, feature selection with Information Gain (IG) is performed to identify the most informative sentiment classification features. This strategy seeks to overcome informal language and noise to improve review context comprehension. The results demonstrated a significant gain in the accuracy of sentiment classification. TFIDF+Bigram+IG achieved 78% accuracy (up 8% from 70% previously), and TFIDF+Trigram+IG achieved 66% accuracy (up 22% from 44% previously). Using this hybrid model, the study significantly enhanced the accuracy of sentiment classification, thereby enhancing the performance of SVM in the face of complex movie evaluations.

1. INTRODUCTION

The majority of the population prefers online movie streaming services, especially movie buffs. These services offer convenience by allowing users to watch various movies from their own homes [1]. Text reviews play an important role in sharing information, where users share their opinions on trending topics, politics, movie reviews, etc. Users can share their opinions and evaluations about movies, thus allowing others to evaluate the quality of the movie based on these reviews [2]. A movie review is an article or piece of content that expresses an individual's opinion regarding a specific film. These reviews contain both positive and negative criticism of the film, which enables the reader to comprehend the film's overall concept and determine whether or not to observe it [3].

NLP-based sentiment analysis employs computational techniques to analyze and interpret the sentiments contained in text documents. This can be accomplished by identifying and categorizing emotions as either positive or negative [4]. In sentiment analysis, classification algorithms can be used to categorize review data into positive or negative sentiment categories. This simplifies data processing and facilitates decision-making based on the reviews' sentiment [5]. In order to distinguish between positive and negative reviews, classification algorithms will learn patterns from these characteristics [6]. In sentiment analysis, naive Bayes, Support Vector Machines (SVM), decision trees, and other machine learning algorithms are frequently employed classification algorithms [7, 8]. Using methods such as feature selection or topic modeling, the dimensionality of features can be reduced during the classification process to identify the most informative and relevant features for distinguishing sentiment [9]. This helps avoid overfitting and optimizes classification performance by using a reduced number of dimensions to predict sentiment while still providing sufficient data [10]. Training and test data sizes for movie reviews are frequently quite substantial [11]. This is due to the abundance of features in the feature space, which can complicate data processing and reduce classification performance [12]. Dimensionality reduction is conducted to solve this issue by removing text document characteristics deemed unimportant. Dimensionality reduction is also useful for optimizing data processing and improving classification performance in sentiment analysis [9, 12].

A key challenge in sentiment analysis of movie reviews is how to classify sentiment in reviews that often use informal language, contain noise and subjectivity, and reflect specific context that reflects movie fans' preferences. This is a complex problem in developing models capable of accurately recognizing sentiment in reviews that are often unstructured and subjective which decreases the performance of the SVM classification model. The SVM problem has difficulties when working with datasets that contain many features. If the number of feature representations in a movie review analysis

dataset is very large, it will be difficult for SVM to obtain complex patterns. SVM uses kernel selection to transfer the data to a higher dimension, where classes can be distinguished with the largest margin. Choosing the optimal kernel can have a great influence on the efficacy of SVM. However, choosing the optimal kernel for sentiment analysis of movie reviews can be difficult, especially if the data characteristics are ambiguous or complex. Class imbalance is a common problem in movie review datasets, indicating that the number of positive and negative reviews may not be equal. SVM may result in unbalanced classification performance if minority classes, such as negative evaluations, are not adequately represented in the model. SVM is a scale-sensitive algorithm. If the features in movie reviews have different scales, SVM may be influenced by features with larger scales, leading to an imbalance in classification.

This research tries to overcome the accuracy problem in sentiment analysis classification of movie reviews by proposing several methods, namely the TF-IDF+N-Gram hybrid model, which is able to extract relevant information from word and phrase sequences, and feature selection with Information Gain (IG), which identifies the most informative and relevant features in sentiment classification so as to improve the algorithm's ability to understand the context and relationship between words in reviews and the SVM algorithm. The selection of these techniques aims to overcome informal language and noise and improve context understanding in reviews. Using this combination, this study achieved a significant improvement in sentiment classification accuracy, strengthening the performance of SVM in the face of complex movie reviews. We propose a new framework to improve the classification performance of the SVM algorithm for movie reviews. This framework generates appropriate features by using the TF-IDF feature weighting method along with the N-Gram model (unigram, bigram, and trigram) [13]. TF-IDF is used to determine the significance of words in the document, while N-Gram allows the retrieval of the context of adjacent words [14]. The combination of these two techniques can result in a more complete and informative representation of movie reviews [5]. Utilize the Information Gain (IG) method to select movie features that are closely associated with positive or negative reviews. Information Gain is a technique that identifies characteristics that significantly contribute to distinguishing emotions [15]. Utilizing IG reduces the number of unimportant features and concentrates on the most informative features. By discarding features that are deemed irrelevant, it reduces the dimensionality of the data and speeds up the classification procedure. After obtaining an appropriate feature representation and reducing the number of features using IG, we also perform SVM model training and evaluation using the processed training data to train the SVM model [16]. Configure the optimal SVM parameters and assess the model's performance using relevant evaluation metrics, such as accuracy, precision, recall, and F1-score.

The contribution of this research is to analyse the effect of using the TFIDF-Ngram method and Information Gain (IG) feature selection by combining the TFIDF-Ngram method involving unigrams, bigrams, and trigrams with feature selection using Information Gain, which can provide valuable insight into the effect of feature relevance in improving the performance of the SVM algorithm for movie review classification. The analysis is conducted to determine the extent to which this combination improves the accuracy or overall performance of the SVM model. because assessing

how much influence feature selection techniques have on the best performance of SVM classification models can provide useful insights in modelling and improve classification accuracy. then evaluate the performance of SVM in the context of sentiment classification for movie reviews. This will provide an understanding of the extent to which SVMs can cope with this classification task and provide guidance for further use of SVMs in sentiment analysis.

2. RELATED WORK

Several previous of research discuss the improved performance of the Support Vector Machine (SVM) algorithm on sentiment analysis film review classification using the TF-IDF feature weighting technique. This study emphasizes the problem of using large features in the film review dataset. The accuracy results obtained are 82.2%. This accuracy gain shows an increase of 11.5% from the previous accuracy gain of 68.7% [12]. This research used several datasets to train and test the SVM algorithm, which resulted in an accuracy of 89.98%. These results show that SVM is a good choice for sentiment classification tasks. Nonetheless, the research suggests that accuracy can be improved by considering more sentence forms. The results of this study have potential applications in improving product sentiment analysis, and future research could consider using more advanced text extraction features. Overall, this research has great potential for understanding and improving sentiment analysis in product reviews [17]. Gini Index-based feature selection method with Support Vector Machine (SVM) classifier is proposed for sentiment classification for a large movie review dataset. The results show that the Gini Index method has better classification performance in terms of lower error rate and accuracy [18]. a linguistic rule-based feature selection method for relevant feature selection in SA that identifies, extracts, and selects the appropriate sentiment features from unstructured datasets and can be further used to classify positive and negative classes. this study proposes an ensemble model in which SVM, Naive Bayes and Random Forest are used as learning algorithms. It is evident from the experimental results that the proposed methodology surpasses the existing baseline methods with 94.7% accuracy in classification [19]. Utilized the particle swarm optimization (PSO) algorithm to improve SVM performance by combining SVM and PSO. In the PSO test, it affects the accuracy of SVM performance so that the accuracy obtained increases from 71.87% to 77% [13]. Proposed a unigram feature extraction technique that aims to break some characters from a string and is then compared with bigram and trigram, unigram+bigram, and unigram+trigram, to improve SVM performance. The accuracy results obtained by SVM with the unigram technique are 84.2%, while the SVM against the bigram technique is 56.2% and the SVM with the trigram technique is 78.8%, which means that the SVM with the unigram technique is superior to other comparison methods [20]. The research utilizes FastText embedding to train word representation on a dataset of over 450,000 tweets. The proposed deep learning model includes convolution, max pooling, dropout, and dense layers with ReLU and sigmoid activations, achieving a remarkable accuracy of 0.925969 on the dataset. The results compare positively against other classifiers, and tweets are found to have informational (54.41%), negative (24.50%), and neutral (21.09%) sentiments related to working from home [21].

3. PROPOSED METHOD

The techniques proposed in this study, namely TFIDF+NGram feature extraction and Information Gain (IG)

feature selection, have the potential to enhance the classification performance of SVM when applied to movie reviews.
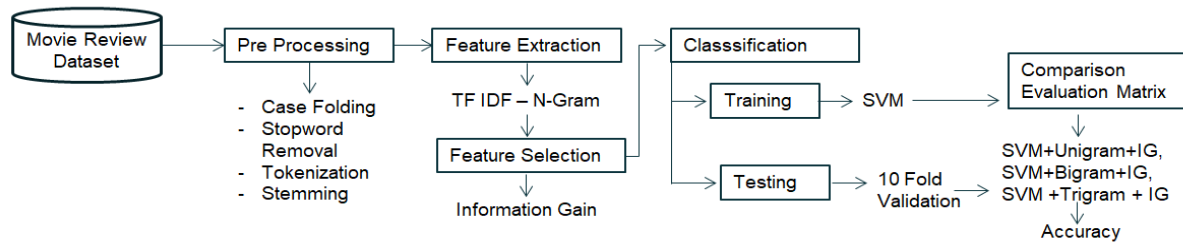

Figure 1. Proposed method for improving the performance of classifiers

Based on Figure 1, the hybrid technique of TFIDF feature extraction with N-Gram approaches (such as unigram, bigram, and trigram) can assist in extracting contextual information and more complex patterns from movie review texts. This method can generate a vector representation that takes into consideration the relative weights of words in the document by considering the frequency of words and word combinations within the review. The system then selects the most pertinent and informative features from the dataset using the Information Gain (IG) technique. This technique measures the amount of information that each feature contributes to the classification. By utilising this feature selection technique, the SVM model is able to prioritise the most pertinent features while ignoring less informative ones, thereby enhancing its performance and efficacy.

The Hybrid TF-IDF, N-Gram, and Information Gain (IG) feature selection model provides a powerful approach for processing and selecting features in sentiment analysis of movie reviews. It allows the model to work with informative words, consider context and word order, and reduce data dimensionality to improve modeling efficiency and performance.

3.1 Data collection

The review film data used in this study was obtained from previous research by Nurdiansyah et al. [22]. This dataset is in the form of review data for Indonesian-language films. The amount of data used was 500, consisting of 250 positives and 250 negatives. This dataset is used to test the performance of the SVM classification algorithm by dividing it into two types, namely training data and test data. The training data used is taken from a collection of reviews of datasets that have been labeled positive and negative. 90% of the entire dataset is used to form a sentiment analysis model. and 10% for test data.

3.2 Data preprocessing

The review film data in general, the stages of data preprocessing used in Natural Language Processing (NLP) are case-folding, this step involves converting uppercase letters to lowercase. For example, "Film Reviews" becomes "film review." Case-folding helps ensure uniformity and consistency in the text data [23]. Tokenization, the process of splitting text or sentences into smaller units, called tokens [1]. Stopword removal involves removing common words such as "the", "and", "in", or "to" that occur frequently in language but often carry little meaningful information in the context of

analysis. This technique helps reduce noise in the data and focuses on more informative terms. [24]. And stemming, process of reducing words to their root or base form. For example, "jumping" and "jumps" would both be stemmed to "jump" [25].

3.3 Feature extraction

In this study, the extraction techniques used are TFIDF and N-Gram, which include (unigram, bigram, trigram), which can be explained as follows:

3.3.1 Term frequency inverse document frequency (TFIDF)

TFIDF is used to calculate the frequency of occurrence of terms in a document [12]. Term frequency (TF) is the frequency with which a term appears in a document. The greater the number of terms that appear, the greater the weight of the document or its suitability value [14]. TFIDF can be formulated in Eq. (1):

$$TF - IDF_{ij} = tf_{ij} \ x \ log \left(\frac{N}{df_i + 1}\right) \qquad (1)$$

IDF counts each word from the total number of documents in the *corpus* with the number of words per document frequency *I* with a value of $IDF_i = log\left(\frac{N}{df_i + 1}\right)$. The use of the TF-IDF method aims to speed up the term calculation process. In addition to speeding up term calculations, TF-IDF can perform efficient weighting and produces accurate results.

Table 1. Example of movie review dataset

Indonesian Text		English Text	
Original Sentences	*film yang bagus dan mendidik*	**Original Sentences**	a good and educational movie
Unigram	*"film", "yang", "bagus", "dan", "mendidik".*	**Unigram**	"a", "good", "and", "educational", "movie".
Bigram	*Bigrams: "film yang", "yang bagus", "bagus dan", "dan mendidik".*	**Bigram**	"a good", "good and", "and educational", "educational movie".
Trigram	*"film yang", "yang bagus", "bagus dan", "dan mendidik".*	**Trigram**	"a good", "good and", "and educational", "educational movie".

3.3.2 Ngram

Defines a method for finding the word set n-gram from a particular document. In the research area, unigrams, bigrams, and trigrams are often used for sentiment analysis. The model N-gram can be explained in Table 1.

Unigram presents the simplest model of the N-gram, consisting of all the individual words in the text. The *bigram model* defines a pair of words that are close together, with each pair of words forming one bigram. Trigrams can be formed in the same way by taking N adjacent words. The N-Gram method is more efficient in providing a better understanding of word position [15].

3.3.3 Feature selection Information Gain (IG)

Feature selection, commonly called variable selection, attribute selection, or feature subset selection, is the process of selecting features that are relevant to the term that is the target of data learning on a problem [26]. Information Gain (IG) can be formulated in Eq. (2):

$$Entropy(S) = \sum_{i=1}^{k}(p_i) log_2(p) \qquad (2)$$

The value of is the proportion of data S with class i and k being the number of classes at the output (S):

$$Entropy(S,A) = \sum_{i=1}^{v}(\frac{sv}{s} \; x \; Entropy(sv)) \qquad (3)$$

The value of v is all possible values of attribute A and is a subset of s where attribute A is worth v. To get the value of Information Gain (IG), use Eq. (4):

$$Gain(S,A) = Entropy \; (S) - Entropy \; (S,A) \qquad (4)$$

Gain value (S,A) is the value of Information Gain, and Entropy (S) is the entropy value before the separator. For comparison, the entropy (S,A) is the entropy value after the separator [27]. An example of a feature set can be seen in Table 2.

Table 2. Feature collection

ID Doc.	"film" "film"	Feature bagus" "good"	"hebat" "great"	Sentiment
D1	Yes	No	Yes	Positive
D2	Yes	No	Yes	Positive
D3	Yes	No	No	Positive
D4	Yes	Yes	Yes	Positive
D5	No	Yes	Yes	Negative
D6	No	No	Yes	Negative
D7	No	Yes	No	Negative
D8	Yes	Yes	No	Negative
D9	Yes	Yes	No	Negative
D10	Yes	Yes	Yes	Negative

Features in Table 2, are examples of snippets of words in a text document that will be weighted, the calculation of feature weights can be shown as follows:

$$Entropy \; (Set) = -\left[\left(\frac{7}{10}\right) log_2 \left(\frac{7}{10}\right) + \left(\frac{3}{10}\right) log_2 \left(\frac{3}{10}\right)\right]$$
$$= 0,97095$$

Value *Entropy positive and negative entropy* that has value *Yes* and *No* on the word "*Film*."

$$Entropy \; (Positive)$$
$$= -\left[\left(\frac{6}{7}\right) log_2 \left(\frac{6}{7}\right) + \left(\frac{1}{7}\right) log_2 \left(\frac{1}{7}\right)\right] = 0,591673$$

$$Entropy \; (negative)$$
$$= -\left[\left(\frac{0}{3}\right) log_2 \left(\frac{0}{3}\right) + \left(\frac{3}{3}\right) log_2 \left(\frac{3}{3}\right)\right] = 0$$

$$Entropy \; (S_{film})$$
$$= \left(\frac{7}{10}\right) 0,591673 + \left(\frac{3}{10}\right) x \; 0 = 0,414171$$

Value *Entropy positive and negative* that has value *Yes* and *No* on the word "*bagus*" is:

$$Entropy \; (positive)$$
$$= -\left[\left(\frac{3}{6}\right) log_2 \left(\frac{3}{6}\right) + \left(\frac{3}{6}\right) log_2 \left(\frac{3}{6}\right)\right] = 1$$

$$Entropy \; (negative)$$
$$= -\left[\left(\frac{3}{4}\right) log_2 \left(\frac{3}{4}\right) + \left(\frac{1}{4}\right) log_2 \left(\frac{1}{4}\right)\right] = 0,811278$$

$$Entropy \; (S_{bagus})$$
$$= \left(\frac{7}{10}\right) 1 + \left(\frac{3}{10}\right) x \; 0,811278 = 0,924511$$

Value *Entropy Entropy positive and negative* that has *value, Yes* and *Not* on the word "*hebat*" is:

$$Entropy \; (positive)$$
$$= -\left[\left(\frac{3}{5}\right) log_2 \left(\frac{3}{5}\right) + \left(\frac{2}{5}\right) log_2 \left(\frac{2}{5}\right)\right] = 0,970951$$

$$Entropy \; (negative)$$
$$= -\left[\left(\frac{3}{5}\right) log_2 \left(\frac{3}{5}\right) + \left(\frac{2}{5}\right) log_2 \left(\frac{2}{5}\right)\right] = 0,970951$$

$$Entropy \; (S_{hebat})$$
$$= \left(\frac{7}{10}\right) 0,970951 + \left(\frac{3}{10}\right) x \; 0,970951 = 0,970951$$

While the Information Gain (IG) value obtained is:

$$Gain(S_{film}) = 0,970951 - 0,414171 = 0,55678$$
$$Gain(S_{bagus}) = 0,970951 - 0,924511 = 0,046439$$
$$Gain(S_{hebat}) = 0,970951 - 0,970951 = 0$$

3.3.4 Classification with Support Vector Machine (SVM)

At this stage, the data review films that have passed preprocessing are classified as having positive or negative sentiment. We propose a feature-weighted model using the frequency inverse document frequency term (TFIDF) and the N-gram model, which includes n=1 (unigrams), n=2 (bigrams), and n=3 (trigrams). N-Gram generates various N-Gram frequencies from the training data to represent the collected documents for the classification process. Extracted features can be selected by using Information Gain (IG) to generate features that are most relevant to movie review data. Our proposed model is then applied to the Support Vector Machine (SVM) classification algorithm.

Support Vector Machine (SVM) is a machine learning algorithm used to classify data by constructing a separating hyperplane between classes of data. SVM has the benefit of handling high-dimensional data and has strong separation ability, and produces good models in generalizing new data. However, SVM is sensitive to unbalanced data and requires

careful parameter tuning. The SVM algorithm is a classification model that requires document text to be converted into vectors or features that are used for classification [28]. The use of the Support Vector Machine (SVM) algorithm as a classification method in sentiment analysis can give good results because SVM is a classification method that often handles the classification of SVM linear models. One of the weaknesses of the SVM algorithm is that it is difficult to apply to large-scale problems; the intended scale is large because the number of samples processed makes it difficult to determine the optimal parameter values. The basic idea of SVM is to find a hyperplane that perfectly separates dimensional data into two classes. However, since the data is not linearly separated, SVM introduces a new kernel-reduced feature space, turning the data into a higher-dimensional separable data space. Usually, dimensional space will cause a lot of computational problems with overfitting [13]. SVM notation is given a linearly separable set of NS={$x_i \varepsilon$ Rn |i=1, 2, …, n}, each point xi has one of the two classes labeled {-1, +1} for=1, 2, …, l, where l is the number of data. It is assumed that both classes -1 and +1 can be completely separated by a hyperplane with the specified dimensions defined:

$$w.x + b = 0 \qquad (5)$$

The pattern x_i, which belongs to class -1 (negative sample), can be formulated as a pattern that satisfies the inequality:

$$w.x_i + b \leq -1 \qquad (6)$$

The pattern, which belongs to class -1 (negative sample), can be formulated as a pattern that satisfies the inequality:

$$w.x_i + b \geq +1 \qquad (7)$$

The largest margin can be found by calculating the value of the distance between the hyperplane and its closest point, which is 1/||w||. For i=1, 2, ..., N, where the operating point (.) is the definition of Eq. (8):

$$\begin{cases} w.x_i + b \geq +1 \ if \ y_i = +1 \\ w.x_i + b \leq -1 \ if \ y_i = -1 \end{cases} for \ i = 1, 2, ..., N \qquad (8)$$

Merging the two equations from Eqns. (6) and (7) produces Eq. (8).

3.3.5 Validation and evaluation

The validation process uses the cross-validation method. In this study, the value of k is set to 10 data subsets. Thus, the dataset is divided into 10 areas. Each piece of data has a different percentage value, which is then evaluated using the confusion matrix method to determine the accuracy of the classification results. We evaluate the performance of the algorithm with a matrix called the Confusion Matrix, where the matrix in each column represents the number of each data point in a predefined class. Each row represents the number of each data point in the predicted class [29].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (9)$$

$$Precision = \frac{TP}{TP + FP} \qquad (10)$$

$$Recall = \frac{TP}{TP + FN} \qquad (11)$$

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad (12)$$

where, true positive (TP) is total identified documents with correct and credible. Whereas false positive (FP) is a lot document fake. True Negative (TN) is total identified documents with true or not credible. False Negative (FN) is total wrong document identified as documents that are not credible. Evaluation using Confusion Matrix method is essential to understand the classification performance by calculating accuracy, precision, recall, and F1 score. The prediction results of the test data are validated by calculating the accuracy using the Confusion Matrix method [30].

4. RESULT AND DISCUSSION

In this experimental stage, we trained the proposed model with a dataset of film reviews in Indonesian, as we have illustrated in Figure 1. We use a machine-learning model with an SVM algorithm. Then apply 10-fold cross validation in the testing process for the entire document data set to measure the performance of classifiers for evaluating the results of accuracy, precision, recall, and f1-score as described in Section 3.6 validation and evaluation.

4.1 SVM experiment results with extraction features TFIDF+NGram

Classification is done using the Support Vector Machine algorithm using a combination of features TFIDF+NGram which includes TFIDF+Unigram, TFIDF+Bigram, and TFIDF+trigram as a feature set. The feature set was tested with three different weighting schemes. First, the features generated are unigrams from the data, and each is tested by weighting TF and weighting TFIDF, then bigrams and trigrams are also tested in the same way. In this test, we conducted a test without applying any feature selection technique to compare SVM performance. The validation process uses 10-fold cross-validation on each weight obtained. The results of the SVM classification with the combination of TFIDF+NGram+IG features can be seen in Table 3.

Table 3. Validation 10 fold SVM with hybrid features

10-Fold Cross Validation	Validation 10 Fold SVM with Hybrid Features		
	TFIDF+Uni gram+IG	TFIDF+Big ram+IG	TFIDF+Trig ram +IG
Fold=1	75%	43,7%	43,7%
Fold=2	78,5%	50%	50%
Fold=3	92,3%	46,1%	53,8%
Fold=4	83,3%	50%	58,3%
Fold=5	75%	66,6%	58,3%
Fold=6	83,3%	50%	58,3%
Fold=7	75%	58,3%	58,3%
Fold=8	91,6%	50%	58,3%
Fold=9	90,9%	72,7%	54,4%
Fold= 0	90,9%	72,7%	54,4%

Based on Table 3, The main finding of the above text is that in classification analysis using Support Vector Machine (SVM)

algorithm with various TFIDF+NGram combination features, TFIDF+Unigram+IG feature set has superior performance compared to other N-Gram models. The average value in the 10-fold validation shows that TFIDF+Unigram+IG has the highest value of 0.836031, while TFIDF+Bigram+IG reaches 0.560358, and TFIDF+Trigram+IG is 0.548354. This means that the use of TFIDF+Unigram+IG features results in higher classification accuracy than other N-Gram features.
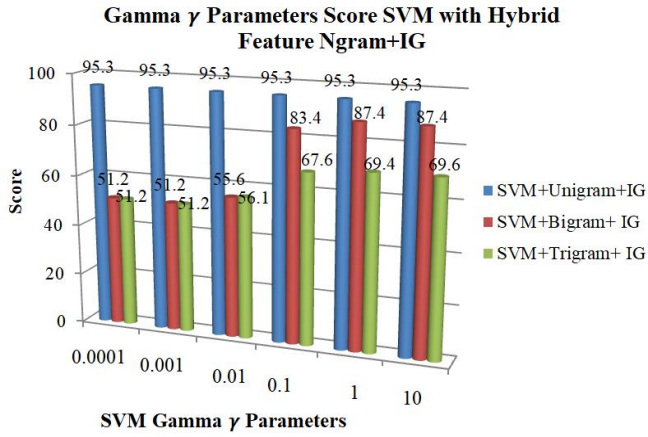
4.2 Analysis of the influence of the SVM kernel

Figure 2. SVM gamma γ parameter scores SVM with hybrid features Ngram+IG

Table 4. Score *gamma γ* by SVM with TFIDF+Ngram+IG

Parameter γ Gamma	Score (%)		
	Average score (%) SVM with Unigram+IG	Average score (%) SVM with Bigram+IG	Average score (%) SVM with Trigram+IG
0.0001	95.3%	51.2%	51.2%
0.001	95.3%	51.2%	51.2%
0.01	95.3%	55.6%	56.10%
0.1	95.3%	83.4%	67.6%
1.0	95.3%	87.4%	69.4%
10.0	95.3%	87.4%	69.6%

Testing and analysis of the classification results are carried out using the gamma parameter on the SVM linear kernel. Manually selecting the gamma (γ) value for the kernel in the SVM model is a good step to optimize model performance. Then train the SVM model by dividing the training data by 90% and 10% testing data. Figure 2 shows the gamma parameter score γ of SVM with Ngram+IG combined features. Repetition of the gamma value is done by applying several predetermined gamma values, namely 0.0001, 0.001, 0.01, 0.1, and 10.0. Table 4 displays the results of the gamma value iteration training on the model used.

Gamma values used include 0.0001, 0.001, 0.01, 0.1, and 10.0. The data used has a data comparison ratio of 90%:10%, of which 90% is training data and 10% is test data. In Table 3, are the results obtained from testing the parameters (gamma), with the highest score of 95.3%, which means SVM+Unigram+IG is superior in the classification process.

4.3 Testing models

The testing process uses the parameters of accuracy, precision, recall, and f-measure to determine the performance of the proposed method and compare the results of SVM performance using the Information Gain (IG) feature selection technique and comparison results. SVM performance without using the Information Gain (IG) feature selection technique. Table 5 shows the results of the SVM classification test with TFIDF+Ngram features, including (unigram, bigram, and trigram). Table 5 describes the results of SVM performance with the TFIDF+Ngram feature without the Information Gain (IG) feature selection technique.

The first test is to test the performance of SVM with the N-gram feature without using the Information Gain (IG) selection feature. The results shown in Table 5 show that the accuracy of the SVM features produced by TFIDF+Unigram is 92%. While SVM uses the TFIDF+Bigram feature which only gets an accuracy value of 70%, and SVM with TFIDF+trigram 44%, we compare the performance of the SVM algorithm using the TFIDF+NGram+IG feature and apply the Information Gain (IG) feature selection technique. The test results can be seen in Table 6.

Table 5. Performance of SVM with hybrid features without selection feature

Algorithm	SVM Classification with Hybrid Feature				
	Accuracy	Precision	Recall	f-Measure	Prediction
TFIDF+Unigram	92%	1.00	0.87	0.93	negative
		0.83	1.00	0.91	positive
TFIDF+Bigram	70%	0.73	0.80	0.76	negative
		0.65	0.55	0.59	positive
TFIDF+Trigram	44%	1.00	0.07	0.12	negative
		0.42	1.00	0.59	positive

Table 6. Performance of SVM with hybrid features and after IG feature selection technique

Algorithm	SVM Classification with Hybrid Feature				
	Accuracy	Precision	Recall	f-Measure	Prediction
TFIDF+Unigram+IG	92%	92%	0.96	0.90	negative
		0.83	0.86	0.95	positive
TFIDF+Bigram	78%	78%	0.77	0.90	negative
		0.65	0.80	0.60	positive
TFIDF+Trigram	66%	66%	0.64	1.00	negative
		0.42	1.00	0.15	positive

In the second experiment, we tested the performance of the algorithm SVM with hybrid feature TFIDF+Ngram and applied the techniques of feature selection Information Gain (IG).

Based on Figure 3, the evaluation of sentiment classification using the parameters precision, recall, and f-measure shows quite good results in the sentiment classification of movie reviews. Precision measures how many of the positive predictions are correct. In this case, the precision value is stable at 1.0, 0.8, and 0.86, while the negative class has a precision score of 0.64, 0.77, and 0.96, which indicates that when the model classifies a review as positive or negative, it is correct. This indicates that the model tends to have little or no error in identifying positive reviews. Recall measures how many of the positive classes were actually identified by the model. In this case, the recall scores for the "negative" class were 1.0, 0.90, and 0.90, while the positive class scored 0.60, 0.15, and 0.95, meaning the model successfully identified all negative reviews. However, the recall for the "positive" class is 0.15, which indicates that the model has difficulty correctly identifying positive reviews. So, the model tends to miss many positive reviews in its classification. F-measure is the combination of precision and recall. The best f-measure is 0.93 for the "negative" class, which shows a good balance between precision and recall in classifying negative reviews. However, the f-measure for the "positive" class decreased to 0.26, which indicates that the model has poor performance in classifying positive reviews.
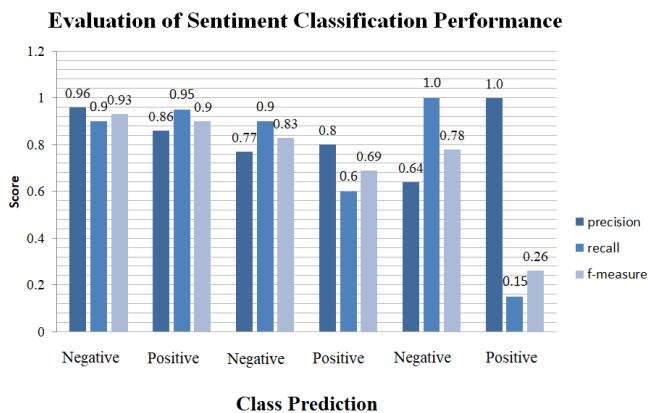

Figure 3. Evaluation of sentiment classification performance

Figure 4 describes the experimental results showing the performance of SVM with hybrid feature TFIDF+Unigram, resulting in an accuracy of 92%. Experiment This showed an increase in SVM performance on the feature TFIDF+Bigram+IG, which obtained an accuracy value of 78% from the previous 70%. Meanwhile, TFIDF+Trigram+IG also experienced an increase in accuracy, namely from 44% to 66%. The comparison graph of accuracy can be shown in 3.

Table 7 presents a comparison of algorithm performance in text classification, including feature extraction and selection, classification techniques, evaluation, and accuracy levels achieved in previous studies. The main finding is that the use of TFIDF+NGram features with feature selection using Information Gain (IG) has successfully improved the performance of the SVM algorithm in text classification, with TFIDF+unigram features achieving the highest accuracy. In addition, this experiment shows improved accuracy on other N-Gram features compared to previous results.

In Table 7, and Figure 5, the performance of classification algorithms using different feature approaches is compared. Research [31], used TF-IDF feature extraction and feature selection using Information Gain (IG). I tested the SVM algorithm with the RBF kernel on the features selected with IG and achieved an accuracy of 87.25%. Other studies [32-35], used feature extraction such as unigram and skip-gram trained with SVM. The results show accuracy between 80% and 90.13%, indicating that the use of unigram and skip-gram feature techniques has good performance in terms of accuracy.
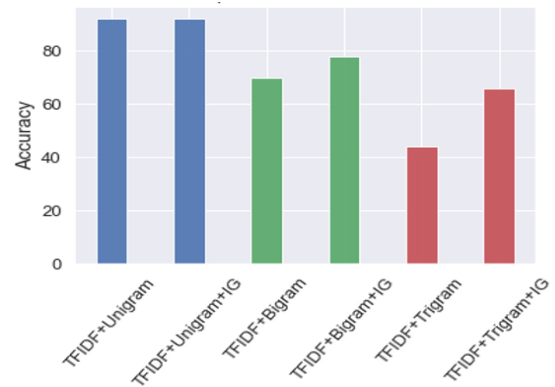

Figure 4. The comparison results of SVM using TFIDF+Ngram feature without SVM TFIDF+Ngram+IG
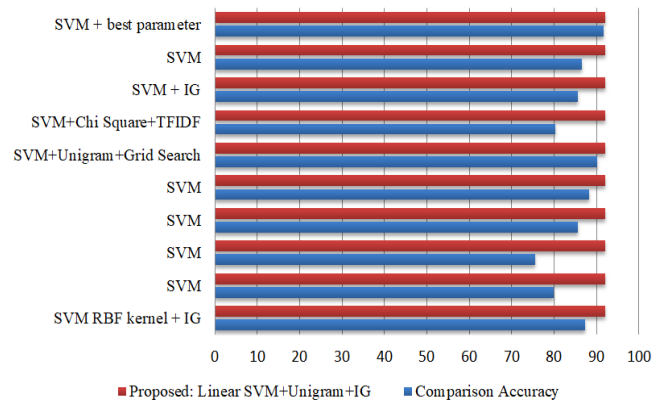

Figure 5. Performance comparison of accuracy

This study proposes a new approach for movie review classification by combining TF-IDF feature extraction and the use of N-grams (unigram, bigram, trigram), as well as feature selection using the Information Gain (IG) algorithm. Support Vector Machine (SVM) classification models with various kernels were used to train this approach. The results show an accuracy of 92%, which is a significant improvement compared to previous studies in movie review classification. This approach has great potential in sentiment analysis and other text classification tasks.

4.4 Analysis of model test results

The application of TFIDF+Ngram fusion technique and Information Gain (IG) feature selection with Support Vector Machine (SVM) algorithm has significantly improved the classification performance. The use of the Bigram feature model (TFIDF+Bigram+IG) Trigram feature model

(TFIDF+Trigram+IG) resulted in a more significant increase in accuracy of 22%, reaching 66% accuracy. resulted in an 8% increase in accuracy compared to the previous model, reaching 78%. The Support Vector Machine (SVM) algorithm proved effective in classifying the polarity of positive and negative reviews. There are several factors that affect accuracy, including random selection of test and training data and computation time.

An important finding is that the combination of TFIDF+Ngram, IG feature selection, and SVM results in a significant improvement in movie review classification performance, with high accuracy achieved. The use of Bigram and Trigram models also showed considerable improvement in accuracy.

Table 7. Performance comparison of classification algorithms based on accuracy

Author's	Feature Extraction or Selection	Classifiers	Evaluation Measure	Best Accuracy
[31]	TF-IDF, IG	SVM RBF kernel + IG	No	87.25%
[32]	TF-IDF, Word2Vec, CBOW and Skip-Gram	SVM	Yes	80%
[2]	TF-IDF, Lexicon based, IG, Correlation, Chi Square, RLPI	SVM	Yes	75.46%
[33]	TF-IDF, Grid-Search for hyper parameter	SVM	Yes	85.65%
[34]	-	SVM	No	88.3%
[35]	Unigram, Grid Search	SVM+Unigram+Grid Search	Yes	9013%
[12]	TFIDF, Chi Square	SVM+Chi Square+TFIDF	Yes	80.2%
[36]	Information Gain (IG)	SVM + IG	Yes	85.65%
[37]	Word2Vector, LDA	SVM	Yes	86.5%
[38]	TF-IDF	SVM + best parameter	Yes	91.63%
Proposed Method	TF-IDF+Ngram+IG	Linear SVM+Unigram+IG	Yes	92%

5. CONCLUSIONS

The conclusion of this research is that the combination of TFIDF+Ngram hybrid technique, Information Gain (IG) feature selection, and the application of Support Vector Machine (SVM) algorithm have a positive impact on the classification performance of movie reviews. The high accuracy, reaching 92%, shows that the model is effective in distinguishing between positive and negative reviews. In addition, the use of Bigram and Trigram features also resulted in a significant improvement in accuracy, with an increase of 8% and 22% respectively. These results reinforce the idea that this approach has great potential for application in sentiment analysis and other text classification tasks. While there are some considerations related to random selection of test and training data and computation time, these findings provide a strong basis for further development in this area.

REFERENCES

[1] Yasen, M., Tedmori, S. (2019). Movies reviews sentiment analysis and classification. In 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), Amman, Jordan, pp. 860-865. https://doi.org/10.1109/JEEIT.2019.8717422

[2] Harish, B.S., Kumar, K., 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

[3] Thet, T.T., Na, J.C., Khoo, C.S. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. Journal of Information Science, 36(6): 823-848. https://doi.org/10.1177/0165551510388123

[4] Kalaivani, P., Shunmuganathan, K.L. (2014). An improved K-nearest-neighbor algorithm using genetic algorithm for sentiment classification. In 2014 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2014], Nagercoil, India, pp. 1647-1651. https://doi.org/10.1109/ICCPCT.2014.7054826

[5] Jagdale, R.S., Shirsat, V.S., Deshmukh, S.N. (2019). Sentiment analysis on product reviews using machine learning techniques. In Cognitive Informatics and Soft Computing. Advances in Intelligent Systems and Computing, pp. 639-647. https://doi.org/10.1007/978-981-13-0617-4_61

[6] Hamzah, M.B. (2021). Classification of movie review sentiment analysis using chi-square and multinomial naïve bayes with adaptive boosting. Journal of Advances in Information Systems and Technology, 3(1): 67-74. https://doi.org/10.15294/jaist.v3i1.49098

[7] Sutriawan, S., Andono, P.N., Muljono, M., Pramunendar, R.A. (2023). Performance evaluation of classification algorithm for movie review sentiment analysis. International Journal of Computing, 7-14. https://doi.org/10.47839/ijc.22.1.2873

[8] Khan, M.T., Durrani, M., Ali, A., Inayat, I., Khalid, S., Khan, K.H. (2016). Sentiment analysis and the complex natural language. Complex Adaptive Systems Modeling, 4(1): 1-19. https://doi.org/10.1186/s40294-016-0016-9

[9] Yulietha, I.M., Faraby, S.A., Widyaningtyas, W.C. (2018). An implementation of support vector machine on sentiment classification of movie reviews. In Journal of Physics: Conference Series, 971(1): 012056. https://doi.org/10.1088/1742-6596/971/1/012056

[10] Liang, H., Sun, X., Sun, Y., Gao, Y. (2017). Text feature extraction based on deep learning: A review. EURASIP Journal on Wireless Communications and Networking, 2017(1): 1-12. https://doi.org/10.1186/s13638-017-0993-1

[11] Gupta, K., Jiwani, N., Afreen, N. (2023). A combined approach of sentimental analysis using machine learning techniques. Revue d'Intelligence Artificielle, 37(1): 1-6.

https://doi.org/10.18280/ria.370101

[12] Larasati, U.I., Muslim, M.A., Arifudin, R., Alamsyah, A. (2019). Improve the accuracy of support vector machine using chi square statistic and term frequency inverse document frequency on movie review sentiment analysis. Scientific Journal of Informatics, 6(1): 138-149. http://journal.unnes.ac.id/nju/index.php/sji

[13] Basari, A.S.H., Hussin, B., Ananta, I.G.P., Zeniarja, J. (2013). Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization. Procedia Engineering, 53: 453-462. https://doi.org/10.1016/j.proeng.2013.02.059

[14] Kim, D., Seo, D., Cho, S., Kang, P. (2019). Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec. Information Sciences, 477: 15-29. https://doi.org/10.1016/j.ins.2018.10.006

[15] Mubaroq, I.M., Setiawan, E.B. (2020). The effect of Information Gain feature selection for hoax identification in twitter using classification method support vector machine. Indonesia Journal on Computing (Indo-JC), 5(2): 107-118. https://doi.org/10.21108/indojc.2020.5.2.499

[16] Gonsalves, T.A. (2020). Feature Selection for Text Classification. Computational Methods of Feature Selection, pp. 273-292. https://doi.org/10.1201/9781584888796-23

[17] Tyagi, E., Sharma, A.K. (2017). Sentiment analysis of product reviews using support vector machine learning algorithm. Indian Journal of Science and Technology, 10(35): 1-9. https://doi.org/10.17485/ijst/2017/v10i35/118965

[18] Manek, A.S., Shenoy, P.D., Mohan, M.C. (2017). Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. World Wide Web, 20: 135-154. https://doi.org/10.1007/s11280-015-0381-x

[19] Saraswathi, N., Rooba, T.S., Chakaravarthi, S. (2023). Improving the accuracy of sentiment analysis using a linguistic rule-based feature selection method in tourism reviews. Measurement: Sensors, 29: 100888. https://doi.org/10.1016/j.measen.2023.100888

[20] Agarwal, B. (2015). One-class support vector machine for sentiment analysis of movie review documents. International Journal of Computer and Information Engineering, 9(12): 2451-2454. http://www.waset.org/publications/10003083

[21] Vohra, A., Garg, R. (2023). Deep learning based sentiment analysis of public perception of working from home through tweets. Journal of Intelligent Information Systems, 60(1): 255-274. https://doi.org/10.1007/s10844-022-00736-2

[22] Nurdiansyah, Y., Bukhori, S., Hidayat, R. (2018). Sentiment analysis system for movie review in Bahasa Indonesia using naive bayes classifier method. In Journal of Physics: Conference Series, 1008(1): 012011. https://doi.org/10.1088/1742-6596/1008/1/012011

[23] Prakoso, B.A., Fanani, A.Z., Riawan, I., Fajri, H., Basuki, R.S., Alzami, F. (2023). Word search with trending reviews on Twitter. Ingénierie des Systèmes d'Information, 28(2): 351-356. https://doi.org/10.18280/isi.280210

[24] Ramasubramanian, C., Ramya, R. (2013). Effective pre-processing activities in text mining using improved

porter's stemming algorithm. International Journal of Advanced Research in Computer and Communication Engineering, 2(12): 4536-4538.

[25] Gogoi, A., Baruah, N., Sarma, S.K., Phukan, R.D. (2021). Improving stemming for Assamese information retrieval. International Journal of Information Technology, 13: 1763-1768. https://doi.org/10.1007/s41870-021-00718-7

[26] Wang, X., Guo, B., Shen, Y., Zhou, C., Duan, X. (2019). Input feature selection method based on feature set equivalence and mutual Information Gain maximization. IEEE Access, 7: 151525-151538. https://doi.org/10.1109/ACCESS.2019.2948095

[27] Kaur, R., Verma, P. (2017). Sentiment analysis of movie reviews: A study of machine learning algorithms with various feature selection methods. International Journal of Computer Sciences and Engineering, 5(9): 113-121. https://doi.org/10.26438/ijcse/v5i9.113121

[28] Haryanto, A.W., Mawardi, E.K. (2018). Influence of word normalization and chi-squared feature selection on Support Vector Machine (SVM) text classification. In 2018 International Seminar on Application for Technology of Information and Communication, pp. 229-233. https://doi.org/10.1109/ISEMANTIC.2018.8549748

[29] Novaković, J.D., Veljović, A., Ilić, S.S., Papić, Ž., Tomović, M. (2017). Evaluation of classification models in machine learning. Theory and Applications of Mathematics & Computer Science, 7(1): 39-46.

[30] Purnomoputra, R.B., Adiwijaya, A., Wisesty, U.N. (2019). Sentiment analysis of movie review using Naïve Bayes method with Gini index feature selection. Journal of Data Science and Its Applications, 2(2): 85-94. https://doi.org/10.34818/JDSA.2019.2.36

[31] Abidin, Z., Destian, W., Umer, R. (2021). Combining support vector machine with radial basis function kernel and Information Gain for sentiment analysis of movie reviews. In Journal of Physics: Conference Series, 1918(4): 042157. https://doi.org/10.1088/1742-6596/1918/4/042157

[32] Lin, S., Zhang, R., Yu, Z., Zhang, N. (2020). Sentiment analysis of movie reviews based on improved word2vec and ensemble learning. In Journal of Physics: Conference Series, 1693(1): 012088. https://doi.org/10.1088/1742-6596/1693/1/012088

[33] Chong, K., Shah, N. (2022). Comparison of Naive Bayes and SVM classification in grid-search hyperparameter tuned and non-hyperparameter tuned healthcare stock market sentiment analysis. International Journal of Advanced Computer Science and Applications, 13(12): 90-94. https://doi.org/10.14569/IJACSA.2022.0131213

[34] Quraishi, A.H. (2020). Performance analysis of machine learning algorithms for Movie Review. International Journal of Computer Applications, 177(36): 7-10. https://doi.org/10.5120/ijca2020919839

[35] Muslim, M.A. (2020). Support Vector Machine (SVM) optimization using grid search and unigram to improve e-commerce review accuracy. Journal of Soft Computing Exploration, 1(1): 8-15. https://doi.org/10.52465/joscex.v1i1.3

[36] 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. In Journal of Physics: Conference Series, 1641(1): 012060.

https://doi.org/10.1088/1742-6596/1641/1/012060

[37] He, H., Zhou, G., Zhao, S. (2022). Exploring E-commerce product experience based on fusion sentiment analysis method. IEEE Access, 10: 110248-110260. https://doi.org/10.1109/ACCESS.2022.3214752

[38] Mustakim, H., Priyanta, S. (2022). Aspect-based sentiment analysis of KAI access reviews using NBC and SVM. IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 16(2): 113-124. https://doi.org/10.22146/ijccs.68903