Improved Accuracy of Sentiment Analysis Movie Review Using Support Vector Machine Based Information Gain

Reza Maulana\textsuperscript{1,}*, Panny Agustia Rahayuningsih\textsuperscript{2}, Windi Irmayani\textsuperscript{3}, Dedi Saputra\textsuperscript{4}, Wanty Eka Jayanti\textsuperscript{5}
\textsuperscript{1,2,3} Sistem Informasi Akuntansi, Universitas Bina Sarana Informatika, Indonesia
\textsuperscript{4,5} Sistem Informasi, Universitas Bina Sarana Informatika, Indonesia
E-mail: reza.rza@bsi.ac.id

Abstract. The quality of a movie can be known from the opinions or reviews of previous audiences. This classification of reviews is grouped into positive opinions and negative opinions. One of the data mining algorithms that are most frequently used in research is the Support Vector Machine because it works well as a method of classifying text but has a very sensitive deficiency in the selection of features. The Information Gain method as feature selection can solve problems faster and more stable convergence levels. After testing on two movie review datasets are Cornell and Stanford datasets. The results obtained on the Cornell dataset is the Support Vector Machine algorithm to produce an accuracy of 83.05%, while for the Support Vector Machine based on Information Gain, the accuracy value is 85.65%. Increased accuracy reached 2.6%. Then, the results obtained on the Stanford dataset is the Support Vector Machine algorithm yields a value of 86.46%, while for the Support Vector Machine based on Information Gain, the accuracy value is 86.62%. Increased accuracy reached 0.166%. Support Vector Machine based Information Gain on the problem of movie review sentiment analysis proved to provide more accurate value.

1. Introduction
A Language is a powerful tool for communicating and conveying information. Language is also a means to express emotions and sentiments. Sentiment analysis is the process of determining whether it is positive, negative or neutral in the contents of a dataset in the form of text (documents, sentences, paragraphs, etc.) [1]. Sentiment analysis in the field of user opinion mining on products, political reviews, movie reviews has now become more popular. Producers and moviemakers through social media and IMDb can find out reviews, views and thoughts from the viewers [2].

Many sites provide a review of a product that can reflect user opinion. One example is the Internet Movie Database (IMDb) site. IMDb is a website that deals with movie and movie production. The information provided by IMDb is very complete, such as who are the actors/actresses who played in the movie, a brief synopsis of the movie, links for movie trailers, release dates for several countries and reviews from other users. When someone wants to buy or watch a movie, other people’s comments and movie ratings usually affect their buying behavior.
There are several classification algorithms widely used for analysis review sentiments include Support Vector Machine, Naïve Bayes, and K-Nearest Neighbour [3]. Some research that has been done in the sentiment classification of online reviews including, Comparative machine learning for the classification of sentiment analysis movie reviews [2]. Sentiment analysis on movie review opinion using the SVM algorithm and PSO [4]. Sentiment classification in online reviews of travel destinations using the NB algorithm, SVM and Character Based N-gram Model [5]. Sentiment analysis of movie reviews and some Amazon.com products using the SVM and Neural Network algorithm [6]. Classification of restaurant review sentiments on the internet uses Cantonese using the NB and SVM algorithm [7].

One of the most frequently used algorithms for data classification is the SVM algorithm. SVM classifies by analyzing data and recognizing patterns, so-called supervised learning methods [4]. SVM has a powerful method for minimizing risk called the regulated linear classification method [8]. Being able to recognize separate hyperplanes that maximize margins between two different classes is also an advantage of the SVM algorithm [9]. However, SVM lacks the problem of selecting appropriate parameters or features [4]. Selecting features in SVM greatly influences the classification accuracy results [10].

Feature selection is important in text classification and greatly affects the classification performance. To enhance the effect of feature selection, many studies try to add optimization algorithms in the feature selection method. Comparative results of feature selection algorithms performed by Chandani [11] between Information Gain, Chi-Square, Forward Selection, Backward Elimination, obtained Information Gain as the best feature selection algorithm.

Thus, SVM classifiers with IG as feature selection will be applied in sentiment analysis in movie reviews. The formulation of the problem in this research is an analysis of improved movie review using the SVM algorithm based IG for feature selection? Then, the purpose of this study is to improve the analysis of movie review analysis using the SVM algorithm based IG for the selection of features needed by users who are used to make decisions in determining movie quality.

The benefits of this research are:

1. The benefits of this research are expected to make it easier for users to make decisions in determining movie quality.
2. Contribute to the development of theories related to sentiment analysis review using the SVM classifier by selecting features using IG to improve achievement.

2. Method

In completing the research, the authors make a framework of thought and proposed methods that are useful as a reference for research so that this research can be done consistently. The problem with this research is that SVM still has weaknesses in the selection of appropriate parameters or features, thus causing a low classification accuracy.

The dataset used in this study is in the form of two movie review datasets, the first Cornell dataset consists of 1,000 positive review data and 1,000 negative review data. And the Stanford dataset consists of 12,500 positive review data and 12,500 negative review data.

Tokenization, Stopwords Removal, and Stemming Methods are the stages of Preprocessing used in research. Information Gain is used as a method of feature selection. Testing uses 10 fold cross validation. Confusion matrix is used to measure the accuracy of the algorithm. Then, the AUC value is measured based on the results of the ROC curve. Experiments using the Weka 3.8 application. To see the framework of thought in detail, see Figure 1 below:

The proposed method is the application of the feature selection method that is IG to improve accuracy in the SVM classifier. SVM algorithm is used because it is very popular and functions well as a text classification method. To see the proposed model in detail, see Figure 2 below:
Figure 1. The Framework of Thought

Figure 2. Proposed Method
3. Results and Discussion
The classification process uses probability calculation values to determine the sentence as a positive or negative class. Sentences are included in the positive class, if the probability value of the sentence for the positive class is greater than the negative class. Meanwhile, the sentence is stated to be included in the negative class, if the positive class probability value is smaller than the negative class. In this study, comparing several classification algorithms namely, SVM, NB, KNN and SVM + IG.

3.1. Comparison of Kernel Functions
In this study, the authors compares some kernel functions to the Support Vector Machine algorithm, to find out the best kernel that can be applied in sentiment review movie analysis. Linear, Polynomial, Radial Basis Function (RBF), and Sigmoid kernels are some of the kernels compared in this study. The following is the result of comparing the accuracy and ROC curves for each kernel, see Table 1 below:

| Kernel   | Accuracy | AUC  |
|----------|----------|------|
| Linear   | 78.45%   | 0.785|
| Polynomial | 81.05%   | 0.811|
| RBF      | 83.05%   | 0.831|
| Sigmoid  | 74.40%   | 0.744|

Based on experiments that have been carried out. Determination of the kernel obtained the best results obtained, namely the highest accuracy value reached 83.05% and the AUC value of 0.831 using RBF Kernel. Therefore, in this study, the RBF Kernel is used in the SVM classification algorithm.

3.2. Parameter Comparison
In this study, the authors conducted an experiment by entering the values of C and epsilon in the Support Vector Machine parameters, to find out the best C and epsilon values that can be applied in the movie review sentiment analysis. The following is the result of comparing the accuracy and ROC curves for each of the C and epsilon values, see Table 2 below:

| Parameter | SVM  |
|-----------|------|
| C         | Epsilon | Accuracy | AUC  |
| 0.0       | 0.0     | 50%      | 0.500|
| 0.0       | 0.5     | 50%      | 0.500|
| 0.5       | 0.0     | 81.35%   | 0.814|
| 0.5       | 0.5     | 81.50%   | 0.815|
| 1.0       | 0.0     | 83.05%   | 0.831|
| 1.0       | 0.5     | 82.55%   | 0.826|

Based on experiments that have been done, the determination of parameters for C and epsilon can affect the accuracy value. The best results obtained are the highest accuracy value reaching
83.05% and the AUC value of 0.831 by determining the value of $C = 1.0$ and $\epsilon = 0.0$. Therefore, in this study, the values of $C = 1.0$ and $\epsilon = 0.0$ are used in the classification of the SVM algorithm.

### 3.3. Comparison of Evaluation and Validation of Results

In this study, testing uses 10 fold cross validation. Confusion matrix is used to measure the accuracy of SVM, NB, KNN, and SVM + IG on the Cornell and Stanford datasets. Then, the AUC value is measured based on the results of the ROC curve. Comparison of accuracy and AUC results can be seen in the Table 3 below:

| Classifier | Dataset Cornell | Dataset Stanford |
|------------|-----------------|------------------|
| SVM        | 83.05%          | 86.46%           |
|            | 0.831           | 0.865            |
| NB         | 80.75%          | 81.28%           |
|            | 0.895           | 0.893            |
| KNN        | 55.95%          | 60.97%           |
|            | 0.572           | 0.626            |
| SVM + IG   | 85.65%          | 86.62%           |
|            | 0.857           | 0.866            |

Comparison chart of the accuracy and comparison chart of the ROC curve values of each algorithm in the Cornell and Stanford datasets. Can be seen in Figure 3 and Figure 4 below:

![Figure 3. Comparison Chart of Accuracy](image1)

![Figure 4. Comparison Chart of ROC Curve](image2)

Based on the above table, it shows that the use of SVM based IG on 10 fold cross-validation testing for movie review on Cornell dataset has the highest accuracy value of 85.65% and AUC value of 0.857 included in the good classification. Then, for the movie review on the Stanford dataset has the highest accuracy value of 86.62% and AUC value of 0.866 included in the good classification.

### 4. Conclusion

In this research, the classification of sentiment analysis of film review was conducted with the classification of SVM. Because the SVM can function properly as a method of text classification. The study used two movie review datasets, the Cornell dataset with 2,000 data and the Stanford
dataset with 25,000 data. From the data processing that has been done, the use of feature selection methods, namely IG, can improve the accuracy of the SVM classifiers. Movie review data can be classified properly into positive reviews and negative reviews.

The accuracy of SVM in Cornell dataset before using the merge with the feature selection method reached 83.05%, with an AUC value of 0.831 included in the good classification after using the merge feature selection method accuracy increased to 85.65%, with an AUC value of 0.857 included in the good classification. Increased accuracy reaches 2.6%. Then, the accuracy of SVM in the Stanford dataset before using the merge with the feature selection method reached 86.46%, with an AUC value of 0.865 included in the good classification after using the merge feature selection method the accuracy increased to 86.62%, with an AUC value of 0.866 included in the good classification. Increased accuracy reaches 0.16%. So the SVM based on IG on the problem of classification analysis of movie review sentiments proved to provide a more accurate accuracy.

To improve this research, the following suggestions are proposed:

1. In subsequent studies, it can use other text classification methods such as Naïve Bayes (NB), K-Nearest Neighbour s (KNN), C4.5 and others.
2. For the next research, it can use other feature selection such as Particle Swarm Optimization (PSO), Genetic Algorithm (GA), Chi-Square and others so that optimal results can be compared.
3. For the next research, it is expected that not only to classify movie reviews but also can use other reviews such as book reviews, online shop reviews and others.
4. The language of the review used is not only English but can use Indonesian or other foreign languages.

References

[1] Kontopoulos E Berberidis C Dergiades T and Bassiliades N, 2013 Expert Systems with Applications Ontology-based sentiment analysis of twitter posts Expert Syst. Appl. 40, 10 p. 4065–4074.
[2] Ahmad E Sazzad M A U Islam M T Azad M Islam S and Ali M H, 2017 Challenges, Comparative Analysis and a Proposed Methodology to Predict Sentiment from Movie Reviews using Machine Learning Big Data Analytics and Computational Intelligence (ICBDAC), 86-91.
[3] Dehkharghani R Mercan H Javeed A and Saygin Y, 2014 Expert Systems with Applications Sentiment causal rule discovery from Twitter Expert Syst. Appl. 41, 10 p. 4950–4958.
[4] Samad A Basari H Hussin B Pramudya I G and Zeniarja J, 2013 Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization Procedia Eng. 53 p. 453–462.
[5] Ye Q Zhang Z and Law R, 2009 Expert Systems with Applications Sentiment classification of online reviews to travel destinations by supervised machine learning approaches Expert Syst. Appl. 36, 3 p. 6527–6535.
[6] Moraes R Valiati J F and Neto W P G, 2013 Expert Systems with Applications Document-level sentiment classification - An empirical comparison between SVM and ANN Expert Syst. Appl. 40, 2 p. 621–633.
[7] Zhang Z Ye Q Zhang Z and Li Y, 2011 Expert Systems with Applications Sentiment classification of Internet restaurant reviews written in Cantonese Expert Syst. Appl. 38, 6 p. 7674–7682.
[8] Weiss S M Indukhyna Nitin and Zhang Tong, 2010 Fundamentals of Predictive Text Mining. London - Springer-Verlag.
[9] Chou J Cheng M Wu Y and Pham A, 2014 Expert Systems with Applications Optimizing parameters of support vector machine using fast messy genetic algorithm for dispute classification Expert Syst. Appl. 41, 8 p. 3955–3964.
[10] Zhao M Fu C Ji L Tang K and Zhou M, 2011 Expert Systems with Applications Feature selection and parameter optimization for support vector machines - A new approach based on genetic algorithm with feature chromosomes Expert Syst. Appl. 38, 5 p. 5197–5204.
[11] Chandani V Wahono R S and Purwanto, 2015 Komparasi Algoritma Klasifikasi Machine Learning Dan Feature Selection pada Analisis Sentimen Review Film J. Intell. Syst. 1, 1 p. 56–60.