Sentiment Analysis on Indonesian Movie Review Using KNN Method With the Implementation of Chi-Square Feature Selection

Imam Prayoga*, Mahendra Dwifebri P, Adiwijaya

Fakultas Informatika, Informatika, Telkom University, Bandung, Indonesia
Email: *prayogaimam@student.telkomuniversity.ac.id, mahendradp@telkomuniversity.ac.id, adiwijaya@telkomuniversity.ac.id

Correspondence Author Email: prayoga.imam2001@gmail.com

Abstract—The advancement and development of the internet is used by the people to support various sectors, one of which is the film industry. Nowadays, people can easily access various movies from available sites. This convenience had led to many reviews about a movie that can be obtained easily. This movie review is very influential on the variety of movies. Freedom of expression on the internet, makes the reviews of a movie vary. For this reason, it is necessary to analyze the sentiment of the movie reviews that are positive or negative. In this research, a sentiment analysis model is build using chi-square selection feature with the KNN algorithm. The final result of this research is able to provide the best classification model with the implementation of stemming. The value of $K = 267$ in selectkbest at the feature selection stage using chi-square, and using the value of $K = 11$ in the KNN parameter. This model produces f1 score value of 86.98%.

Keywords: Movie Review; Sentiment Analysis; Chi-Square; KNN

1. INTRODUCTION

Along with technological advancements, the process of finding out for information is also evolving. One of the information that frequently search by people is the information about film industry. The film industry is an enterprise that each and every year is growing well[1]. It can be effected by a great enlarge of internet users around the world[2]. Internet gives an easy access to watch any kind of movies from any websites which provided by it. By that easy access, people need to knowing the fine of the movie. Nowadays, lot of sites provide the information about movie review[3]. With the availability of existing movie reviews data, lot of studies have been conducted, one of the Natural Language Processing that is relevant to movie review data is sentiment analysis.

Sentiment analysis is the process of extraction and classification based on the sentiment from a text[4]. This analysis is an important information on the development of any products, one of them is movie production[5]. The result of the sentiment analysis of movie review will classify the review be a positive review or a negative review. Positive review point out that the comment given has a good review, it means that the film reviewed is a good movie. However, negative review indicates that the comment given have poor grades, or in other words, the movie that reviewed is not a good movie.

One of the issues in sentiment analysis is the number of the features. A large number of features can reduce classification performance[5]. For this reason, a feature selection process is needed. There are several feature selection techniques that often used in sentiment analysis. In this study, the feature selection that will be used is Chi-Square technique to calculate the degree of a feature on a class. Paper [6] which conducted by Nurhayati, et al. in 2019, chi-square method proved in improving the accuracy of classification performance from 73.33% to 93.33%. The method used in the study [6] is Naïve Bayes.

In this study, the method to be used is K-Nearest Neighbor. This method is chosen based on the previous research. In [7] research by H. Jelodar, et al. in 2021, KNN method is considered a high-quality approach in behavior analysis, especially in sentiment analysis. Paper [8] was conducted by N. Daeli, A. Adiwijaya in 2020 used KNN method in sentiment analysis of movie review and executed an accuracy of 96.8%. In addition, this study analyzes sentiment from Indonesian movie review datasets and implement chi-square to perform feature selection. This study uses several references from previous studies that are related to sentiment analysis, classification, selection feature, preprocessing and evaluation. Sentiment analysis is frequently used to find out product reviews, apps, films or other data.

In paper [5] by M. Hamzah in 2021, discussed the classification of movie review sentiment but uses the multinomial naïve bayes method with adaptive boosting using chi-square selection feature. The study examined the addition of adaptive boosting and the effect of chi-square selection feature, 85.74% was once acquired by only using chi-square selection feature without adaptive boosting and 81.39% was obtained without using adaptive boosting and chi-square selection feature.

Research [9] which conducted by F. Chandra, Y. Sibaroni and 2019 used TF-IDF as extraction feature and chi-square as selection feature in sentiment analysis on book reviews using Support Vector Machine method. The results of the study were shown through a confusion matrix and compared the performance of the method on two kernel, Linear kernel and Gaussian RBF. The results bought 70.7% on linear kernel with TF-IDF extraction feature and chi-square selection feature. Meanwhile, an accuracy result of 74.2% was obtained on Gaussian RBF kernel with the same extraction feature and selection feature.

Other research was conducted by R. Sari in 2020 [10]. This research discussed performance of KNN method on sentiment of tourist attraction reviews. The preprocessing process used are tokenization, filter stopword and
also N-Gram. This paper only focuses on the performance of KNN method without using any feature selections. Then checking out was carried out on K values from 1 to 10. As the result, K=7 became the best K with an accuracy score 77.01%, precision 92.39%, recall 61.56% and AUC score 0.894.

In research [8] measured the accuracy of KNN method towards movie review using information gain selection feature. The research examined the impact of information gain selection feature on the performance of KNN method in conducting sentiment analysis with thresholds for information gain values divided into 5, it’s 0.1, 0.2, 0.3, 0.4 and 0.5. Then the fine K value will be searched by looking at to the best average value taken from each threshold information gain value. In this study, the highest average accuracy is K = 3 with an average accuracy of 83.45%. After that, the performance of KNN method with information gain selection feature will be compared to KNN method without information gain. As a result, without feature selection, KNN method only executed an accuracy of 60%. Meanwhile, with feature selection, KNN achieved an accuracy of 96.8%.

Other research that has been conducted by B. Jonathan, et al. in 2019 [11]. This research conducted a sentiment analysis of restaurant reviews using random forest method. Data collected as many as 150,000 reviews with three labels, those are positive that have a rating above three, negative with reviews that have a rating below three and neutral that have a rating of three. In this research, preprocessing process used are lowercasing, tokenization, remove punctuation, stopwords removal, pos tagging and lemmatization. Then it used TF-IDF as feature extraction. This study divided the data to 80% training data and 20% testing data. The result acquired from this study achieved an accuracy of 92%.

2. RESEARCH METHODOLOGY

In this research, the system configuration is sentiment analysis by K-Nearest Neighbor and chi-Square. The flow of this system model is illustrated in Figure 1:

![Figure 1. System Design](image)

2.1 Dataset

The dataset used in this research is movie reviews of Indonesian-language review. The dataset was obtained from wovfreak.blogspot.com website, there are 5,529 film review data and has been labeled with positive label as much as 2,884 datas and negative class as much as 2,645. When data preparation process was executed, it has lot of duplicate data, so the step of drop duplicates column must be completed. The data have decreased in size to 4,208 with a distribution of 1,333 on negative label and 2,857 on positive label.

![Figure 2. Distribution of Number of Datasets Labels](image)
2.2 Preprocessing

When all the data is collected, the next step of this research is preprocessing. Preprocessing is the process of change the data format to be an easier format and effective to be processed, so when we go to the classification process, the result will be more accurate. In this research, the preprocessing step is divided into six processes, they are cleansing, case folding, normalization, tokenization, stopword removal and stemming.

Cleansing is a process to remove any components that not effect to classification process. Commonly, data tends to be unstructured and often very noise, cluttered/disorganized[12]. The components that removed are punctuation, double space, emoji, html format, url format, number and symbol.

Case folding is process to change all the letters into the lowercase letters. This step is used to make the data structured and consistent. For example in the sentence “Menit awal ngacak” to “menit awal ngacak”.

Tokenization is process to split sentence into tokens based on space as needed of the system[13]. For example in the sentence “menurut saya ini film mengecewakan” to “[menurut], [saya], [ini], [film], [mengecewakan]”. Stopword removal is process to remove words that have no effect on classification process[13]. List of the words in stopword is word that commonly used in daily. The input data for this process is the result of tokenization process. For example input “[menurut], [saya], [ini], [film], [mengecewakan]” to “[film], [mengecewakan]”.

Stemming is a process to change word into the root word. This process removes the affixes of the word, they are suffixes, prefixes and the combination of both[13]. For example, the sentence “[film], [mengecewakan]” to “[film], [kecewa]”.

2.3 TF-IDF

After preprocessing process is done, the next step is to calculate the weight of the word used. This research uses TF-IDF weighting for the feature extraction. Weighting is the process to convert input words into numerical data. TF-IDF works for calculating the frequency of occurrence of a word in the data[14]. The formula of TF-IDF word weighting can be seen from the following equation:

\[
 w(t, c) = tf(t, d) \times idf(t) = tf(t, d) = \log \frac{N}{df(t)}
\]

Description:

\( w(t, c) \) = term(t) weight in document(d)
\( tf(t, d) \) = number of occurrences of term(t) in document(d)
\( idf(t) \) = number of documents frequency inverses each word
\( df(t) \) = number of documents frequencies each word
\( N \) = total number of documents

2.4 Chi-Square

The next step after feature extraction stage is feature selection using chi-square method. Feature selection is a word selection process carried out to reduce irrelevant features before the classification process. The use of Chi-Square feature selection has proven effective in several previous studies[5][9]. Chi-Square is one of the feature selection methods that works by calculating the relationship or the degree of dependence between an existing term(t) and the class[15]. After that, each term value will be sorted from the highest. The formula for Chi-Square Feature selection formula, as follows [15] :

\[
 X^2(t, c) = \frac{N(A \times D - B \times C)^2}{(A + B) \times (C + D) \times (A \times C) \times (B + D)}
\]

Description:

\( X^2(t, c) \) = term(t) searched for in class c
\( N \) = amount of data
\( A \) = amount of class C documents containing term(t)
\( B \) = amount of data containing term(t) but not in class C
\( C \) = amount of data in class c that is found but do not contain term(t)
\( D \) = amount of data not included in class C data and do not contain term(t)

In using Chi-Square, it is necessary to know the single chi-square value of a particular word. The formula for the single chi-square value of each word is, as follows:

\[
 X^2(t) = \sum_{c=1}^{C} X^2(t, c)
\]
2.5 Split Data

This research divided the data into train data and test data. 80% data is allocated to data train, meanwhile 20% of the rest of the data is allocated to data test. The result of the splitting data can be seen in the following table:

<table>
<thead>
<tr>
<th>Split Data</th>
<th>Data Train</th>
<th>Data Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Data</td>
<td>3366</td>
<td>842</td>
</tr>
<tr>
<td>Positive label</td>
<td>2273</td>
<td>602</td>
</tr>
<tr>
<td>Negative label</td>
<td>1093</td>
<td>240</td>
</tr>
</tbody>
</table>

2.6 K-Nearest Neighbor

K-Nearest Neighbor (KNN) is a classification method for an object on train data taken from its k nearest neighbors. The KNN method has three advantages, namely easy to understand, good performance, and easy to implement parameter tuning to adapt to the needs of research to achieve a better results[16]. Because this method is a classification based on distance calculation, KNN will calculate the distance of one test data with all the train data using Euclidean distance[8]. The formula of Euclidean distance is, as follows:

\[ D(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \]  

Description :
\( D \) : distance between points  
\( x \) : data training  
\( y \) : data testing

2.7 Evaluation

Evaluation aims to measure the performance of the system model with a confusion matrix. Confusion matrix is used to see the actual information and the prediction result of a classification system. The performance of a classification system commonly evaluated by matrix data. The display of confusion matrix table can be seen from the following table:

<table>
<thead>
<tr>
<th>Name</th>
<th>Actual Values</th>
<th>Predicted Values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td></td>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td></td>
<td>FN</td>
<td>TN</td>
</tr>
</tbody>
</table>

Description:
\( TP = \) both a positive real value and a positive forecasted value  
\( FN = \) a negative forecasted value and a positive real value  
\( FP = \) a negative real value and a positive forecasted value  
\( TN = \) both a negative real value and a negative forecasted value

To measure the evaluation of performance from classification system, the confusion matrix calculates the precision, recall and f1-score[17]. The formula of calculating evaluations can be seen from the following formula:

\[ Precision = \frac{TP}{TP+FP} \]  

\[ Recall = \frac{TP}{TP+FN} \]  

\( F1 \) score is the harmonic mean of precision and recall[18]. The formula of \( F1 \) score, as follows:

\[ F1\text{score} = \frac{2 \times (precision \times recall)}{precision + recall} \]
3. RESULT AND DISCUSSION

In this study, 4208 movie review datasets were used from wovfreak.blogspot.com website and divided into train data and test data with distribution ratio of 80:20, with the total of train data is 3366 datasets, and 842 datasets in total for test data. There are three scenarios conducted in this study. The first scenario tests the optimal K on selectkbest in feature selection stage. The second scenario tests the preprocessing stage by looking at the effect of using stemming phase. The third scenario tests the effect of K values on KNN parameters.

3.1 Comparison of k values in SelectKBest to KNN Performance

In the first scenario, a test was conducted to see the effect of the k value on selectkbest in the feature selection stage. The K value tested starts from 50 to 10000 with a multiple of 50. The KNN algorithm is used as the classifier in this scenario. In this scenario, stemming is used in the preprocessing stage, and the K value in the KNN parameter is set to 5.

![Performance Graph of K value on selectkbest](image)

Based on the resulting graph, the greater value of K used, the resulting f1 score decreases. In previous paper [19] that conducted by D. Ratmana, et al. in 2020 also examined the effect of the k value in selectkbest on the performance of the classification method. The results from that research show that for the KNN method, the smaller k value used, the more f1 score of KNN increases. This is directly proportional to this research. The results of precision, recall and f1-score, each has a different k value for the highest score, the results can be seen based on the following table:

<table>
<thead>
<tr>
<th>k value</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>150</td>
<td>84.33%</td>
<td>88.53%</td>
<td>86.38%</td>
</tr>
<tr>
<td>250</td>
<td>82.90%</td>
<td>91.02%</td>
<td>86.77%</td>
</tr>
<tr>
<td>1250</td>
<td>72.26%</td>
<td>98.90%</td>
<td>83.90%</td>
</tr>
</tbody>
</table>

This scenario is continued by looking in more detail at the value of k in the range of 1 to 1000, this is done because based on the resulting graph, the value of k is quite stable at the beginning. The purpose of this analysis is to find the most optimal k value for the best f1 score.

<table>
<thead>
<tr>
<th>k value</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>267</td>
<td>82.12%</td>
<td>92.35%</td>
<td>86.94%</td>
</tr>
</tbody>
</table>

Based on the result, the highest f1 score is obtained at a value of k = 267.

3.2 The Effect of Stemming

In scenario 1 shows that the best k to improve the performance of the model in this study is k = 267. Scenario 2 tests the effect of stemming stage on the performance of the KNN method. In this scenario, the feature selection stage uses k = 267 as the k value in selectkbest, and the K value in the KNN parameter is set to 5.

<table>
<thead>
<tr>
<th>Preprocessing</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stemming</td>
<td>82.12%</td>
<td>92.35%</td>
<td>86.94%</td>
</tr>
<tr>
<td>No Stemming</td>
<td>80.64%</td>
<td>91.36%</td>
<td>85.66%</td>
</tr>
</tbody>
</table>
Based on the results from table 5, the highest precision at 82.12%, recall at 92.35% and f1 score at 86.94%. All of these scores are obtained by involving stemming in the preprocessing stage. While in the experiment that does not involve stemming, there is a decrease in precision by 1.48%, recall by 0.99% and f1 score by 1.28%.

Paper [20] was conducted by A. Pradana, M. Hayati in 2019 also shows that the use of stemming in preprocessing stage can improve the performance of the model. In KNN method, feature that are not very relevant can interfere with the performance of the model. This proves that the use of stemming is useful because it can reduce the number of features by cutting the affix of word.

3.3 The Effect of K values on KNN Parameters

In this scenario, a model performance test is conducted by looking at the effect of the k value with a range of 1 to 30 on the k nearest neighbor parameter on its performance. This experiment was conducted using the chi-square feature selection with a value of k = 267 in selectkbest, and stemming was used in the preprocessing stage.

![Figure 5. Performance Graph of K value on Metrics F1 Score](Image)

In Figure 5, shows the performance graph of K values on KNN parameter, the highest graphic value is achieved using the value of K=11 with f1 score of 86.98%. The performance graph shows that a larger value of K shows better f1 score that a smaller number of K. This is evidenced by the graph which tends to rise of the k value with a range of 1 to 30. Previous research that has been conducted by A. Pamuji in 2021 [21] also compares the value of K in the range of 1 to 18 against the performance of the KNN model, and the best value of K is achieved at K = 17. It can be concluded that the K value affects the performance of the KNN model.

4. CONCLUSION

Based on the experiments and scenario tests that have been carried out, researchers have constructed a system that can analyze sentiment analysis on Indonesian movie review using KNN method and the implement of chi-square as the feature selection, with the first experiment conducted was to compare the K values for selectkbest in feature selection stage within a range of 1 to 10000, the second experiment looks at the effect of using stemming in preprocessing stage, and the third experiment conducted was to look at the effect of K values on KNN parameters on model performance. From the results of the experiments that have been carried out, researchers concluded that the value of K in selectkbest at the feature selection stage affects the performance of the model, where the value of K tends to be low, resulting in better model performance. Then the use of stemming in the preprocessing stage, resulting in more optimal model performance than stages that do not involve stemming. In the classification stages using KNN, the value of K=11 is the most optimal value for its performance. The best model obtained in this study is by using a stemming dataset. The value of k on selectkbest in the feature selection stage is 267, and the K=11 value in the KNN algorithm managed to get f1 score of 86.98%. In this study, the number of datasets is only 4208 with 2875 positive labels and 1333 negative labels. The preprocessing stages does not involve the normalization of the distribution of the labels. Adding the normalization stage can also be used to analyze its effect on model performance. Try to analyze other parameters in KNN such as weights, algorithm, p and metric. The use of hyperparameters tuning is also highly recommended to be done in order to find the most effective parameters. In addition, researcher also suggest to compare KNN method with other classification algorithm.

REFERENCES


