**Building of Informatics, Technology and Science (BITS)**
Volume 5, No 1, June 2023 Page: 162−170
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i1.3595

# Sentiment Analysis on Movie Review from Rotten Tomatoes Using Logistic Regression and Information Gain Feature Selection

**Arsenio Jusuf Abimanyu*, Mahendra Dwifebri, Widi Astuti**

Faculty of Informatic, Informatic, Telkom University, Bandung, Indonesia
Email: [1,*]arsenioabi@student.telkomuniversity.ac.id, [2]mahendradp@telkomuniversity.ac.id, [3]widiwdu@telkomuniversity.ac.id
Correspondence Author Email: abi2001abii@gmail.com
Submitted: **07/06/2023**; Accepted: **25/06/2023**; Published**: 29/06/2023**

**Abstract**−The current state of technology can positively affect how the internet is used as well as how information, particularly information on the film industry, is disseminated online. Many movie reviews can be simply found thanks to this ease. Reviews of movies have a big impact in the various ways movies are available. Thanks to the ease of various information on the internet, the number of movie reviews has become diverse. Consequently, conducting a sentiment analysis is required. In this research, classification method used is Logistic Regression. The method was chosen because it has accurate classification accuracy. In this study, Information Gain was also chosen as a feature selection because it is good enough to do a filter approach in classification. Furthermore, for feature extraction, TF-IDF was chosen because it can overcome data imbalance in the dataset. The best model resulting from this research is a model built without using stemming in the preprocessing stage, without using information gain feature selection, and using parameters in Logistic Regression which produces an f1-score of 76.50%.

**Keywords**: Movie Review; Sentiment Analysis; Information Gain; Logistic Regression

## 1. INTRODUCTION

Along with contemporary technical developments, the act of finding knowledge is expanding fairly quickly. One of them is information about the film industry, which can have a major influence on the film industry. The movie industry is one of the cultural industry fields with a wide market around the world [1]. It is undeniable that the internet is a major factor in the dissemination of information about cinema in the world. The internet can also penetrate the dimensions of life, time, and even the space of its users. The Internet also provides access to a variety of resources, both research results and articles from research in various fields [2]. With internet services, people can give and receive movie reviews. Currently, there are many websites that provide information about movie reviews [3]. If movie review data can be processed well enough, credible information about movie reviews can be obtained. One way to find out is by using Sentiment Analysis. A science known as sentiment analysis is used to evaluate the positive or negative values of ideas, attitudes, products, organizations, events, and so forth [4]. Sentiment analysis is very important and also aims to further develop the product in the future, an example is Rotten Tomatoes movie reviews. Ratings on Rotten Tomatoes are based on an average of reviews from highly subjective professional movie critics.

Research on the classification of movie review sentiment using KNN method and Information Gain as a selection feature conducted by Ria Ine Pristiyanti in 2018 [5] tested adding a high number of term variations based on the Information Gain value threshold. Highest accuracy value of 92% was obtained by using optimal k = 5 value in the KNN classification process. In this study, researchers want to use the Information Gain selection feature which has previously proven effective for selecting features on movie review datasets but with a different classification method, namely Logistic Regression and using TF-IDF as an extraction feature and try to do it using Information Gain and without using Information Gain in this study.

The extraction feature that will be used in this research is TF-IDF which has previously been used for sentiment analysis research on English book reviews by Chandra Gilang Kencana in 2019 [6] which obtained an accuracy value of 74.2% using the SVM method. The study used TF-IDF as an extraction feature and Chi Square as a selection feature. The confusion matrix in the study shows a comparison of the performance of the method on both kernels, namely Linear kernel and the Gaussian RBF kernel. Result of the accuracy obtained 70.7% on the use of Linear kernel with TF-IDF weighting and Chi-Square feature selection. While the Gaussian RBF kernel produces an accuracy of 74.2% using TF-IDF feature weighting and Chi-Square feature selection, TF-IDF extraction feature was chosen because it has proven effective and has succeeded in getting good and optimal results in previous studies.

There is also other research on this topic from Priyanka H S in 2019 [7] discussing the classification of sentiment analysis on movie reviews using the Logistic Regression method with Bi-gram as an extraction feature. The study used binomial numbers, namely 0 in Logistic Regression, showing results of Precision 88%, Recall 88%, and F1-Score 88%. At binomial number 1 for Logistic Regression results of Precision 87%, Recall 89%, and F1-Score 88%. Meanwhile, for Logistic Regression which is taken on average from the dataset to be used as a Bi-gram, it gets an accuracy of 88% , these excellent results make researchers want to make a comparison using the default parameters in Logistic Regression and parameters that can be changed in Logistic Regression.

Amelia Syahadati in 2021 [8] discussed the sentiment analysis of the implementation of the PSBB in DKI Jakarta and its impact on the JCI movement from opinions on Twitter using a comparison of the Logistic Regression, KNN, Random Forest, and Naïve Bayes methods. Study shows a comparison of the final results of accuracy which all have the same model classification accuracy rate, namely for prediction on observation has an average of 75%. Then for the comparison accuracy of the classification method, the Logistic Regression method has an accuracy of

**Building of Informatics, Technology and Science (BITS)**
Volume 5, No 1, June 2023 Page: 162−170
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i1.3595

75%, the KNN method has an accuracy of 70%, the Random Forest method has an accuracy of 60%, and finally Naïve Bayes method has an accuracy of 70%. It can be concluded that Logistic Regression method is right method in analyzing the sentiment.

Research conducted by Bern Jonathan in 2019 [9] describes the sentiment analysis of customer reviews at one of the restaurants on Zomato using the Random Forest method. Collected data with a total of 150,000 multi-label reviews, the first label is a positive label with reviews that have a rating above three, then a negative label on the second label with reviews that have a rating below three. In the study, the preprocessing process carried out included lowercasting, tokenization, remove punctuation, stopword removal, pos tagging, and lemmanitazion. Then, feature extraction is done using TF-IDF. The research divided the data into 80% for train data and 20% for test data. Final results obtained from the study achieved an accuracy of 92%. Positive recall is good with 99% accuracy while neutral has the best precision at 96%.

The best f1-score value of 76.50% was achieved in this study's trials after a method using Logistic Regression classification, TF-IDF for feature extraction, and Information Gain for feature selection was developed.

## 2. RESEARCH METHODOLOGY

System developed in this research uses a classification technique called logistic regression to model sentiment analysis of English movie reviews, Information Gain as feature selection, and also TF-IDF as feature extraction. Figure 1 shows how the design flow of this system.

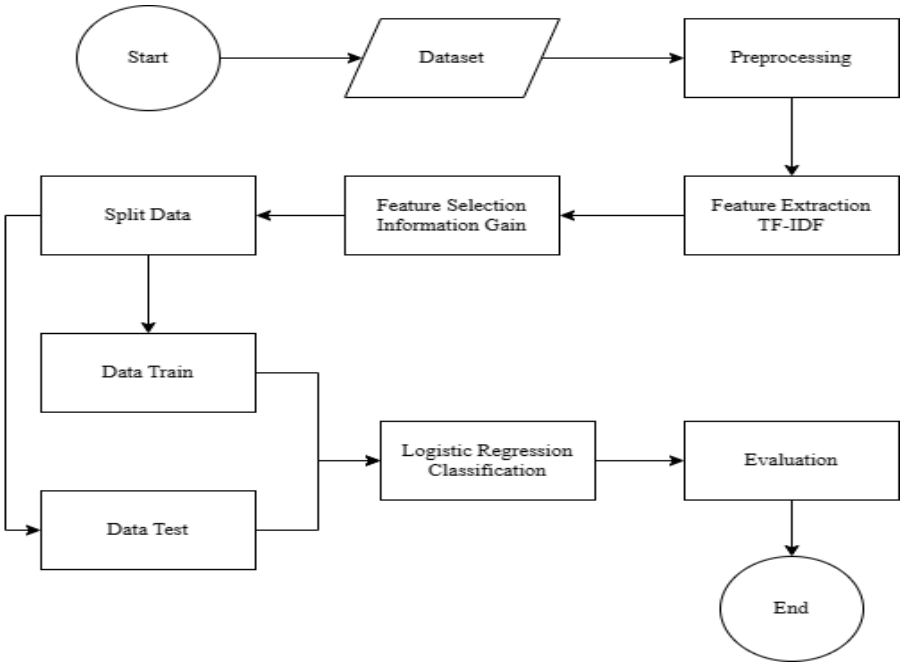


**Figure 1.** System Design Flow

### 2.1 Dataset

The dataset used is English-language movie review data taken from the Rotten Tomatoes website through Kaggle with a total of 1,130,017 data. The dataset is then sorted into reviews starting from 2020 and above with a total of 45,746 data, this is done to keep the sentiment relevant and cut the running time which is quite long. There are two classes in this data label, Fresh for the positive class contains 28,986 data and Rotten for the negative class contains 11,751 data, the total net data is 40,737. The following are examples of datasets that have been labeled as shown in Table 1.

**Table 1.** Data Example

| Label | Sentence |
|---|---|
| Fresh (Positive) | "Betty Davis gives the finest performance of her career. If there was ever any doubt in your mind that she was one of the greatest of screen actresses, her performance here will dissipate ail doubt." |
| Rotten (Negative) | "It feels as if it's more about how a couple can playfully fight while facing off in a courtroom rather than whether or not women should be treated impartially." |

In the label distribution seen in Figure 2, it shows that the distribution of label values is not balanced on labels 1 (positive) and 2 (negative) taken in the label column called review_type. The positive class contains 28,986 data, while the negative class contains 11,751 data. The total net data from the distribution of the two labels is 40,737.

**Building of Informatics, Technology and Science (BITS)**
Volume 5, No 1, June 2023 Page: 162−170
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i1.3595

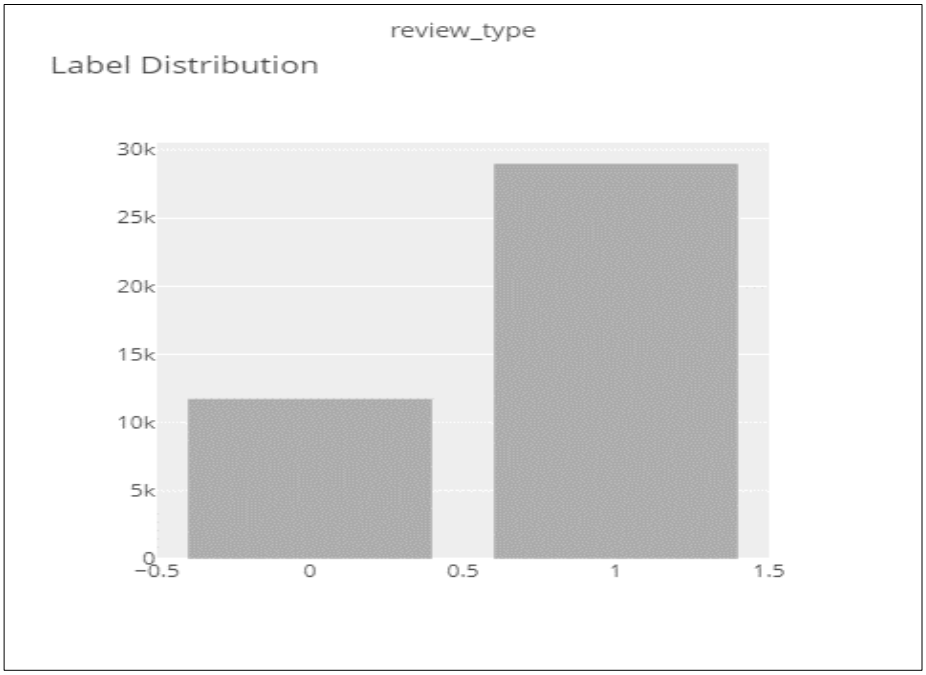


**Figure 2.** Dataset Label Count Distribution

### 2.2 Preprocessing

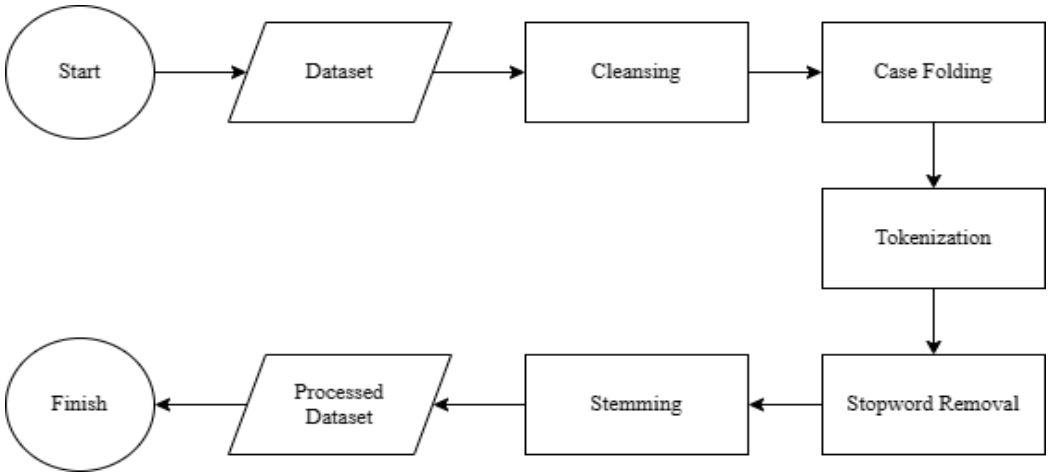


**Figure 3.** Preprocessing Stages

Preprocessing is the next stage after all the data has been collected and prepared. Preprocessing is a mandatory stage where the data will be cleaned first before the classification process is performed. The main purpose of preprocessing is to improve data quality by removing unwanted words and organizing raw data into a format that can be processed and analyzed further, the advantage is the reduction of feature size space. Data sorting starting from 2020 and after is done as preprocessing in this study, changing the data type in the review_type column from the original class name Fresh to 1 while Rotten to 0, Cleansing, Case Folding, Tokenization, Stopword Removal, and Stemming. Preprocessing steps carried out in this study can be seen in Figure 3.

#### 2.2.1 Cleansing

Cleansing is the step of removing considered elements such as noise [10]. Cleansing is the first stage carried out in this research at the preprocessing stage so that the preprocessing process runs better, improves data quality, and ensures the consistency of the data. Table 2 provides a Cleansing sample sentence.

**Table 2.** Cleansing

| Sentence | Cleansing results |
|---|---|
| “Betty Davis gives the finest performance of her career. If there was ever any doubt in your mind that she was one of the greatest of screen actresses, her performance here will dissipate ail doubt.” | “Betty Davis gives the finest performance of her career If there was ever any doubt in your mind that she was one of the greatest of screen actresses her performance here will dissipate ail doubt” |

**Building of Informatics, Technology and Science (BITS)**
Volume 5, No 1, June 2023 Page: 162−170
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i1.3595

### 2.2.2 Case Folding

Case Folding is a work step that changes the letters in capitalization (a-z) [11]. Table 3 shows an illustration of a Case Folding sentence.

**Table 3.** Case Folding

| Cleansing results | Case Folding results |
|---|---|
| "Betty Davis gives the finest performance of her career If there was ever any doubt in your mind that she was one of the greatest of screen actresses her performance here will dissipate ail doubt" | "betty davis gives the finest performance of her career if there was ever any doubt in your mind that she was one of the greatest of screen actresses her performance here will dissipate ail doubt" |

### 2.2.3 Tokenization

Tokenization is an important part of text data preprocessing which is the process of breaking or slicing a sentence, document and graph into tokens [11]. Table 4 provides an illustration of Tokenization.

**Table 4.** Tokenization

| Case Folding results | Tokenization results |
|---|---|
| "betty davis gives the finest performance of her career if there was ever any doubt in your mind that she was one of the greatest of screen actresses her performance here will dissipate ail doubt" | ["betty", "davis", "gives", "the", "finest", "performance", "of", "her", "career", "if", "there", "was", "ever", "any", "doubt", "in", "your", "mind", "that", "she", "was", "one", "of", "the", "greatest", "of", "screen", "actresses", "her", "performance", "here", "will", "dissipate", "ail", "doubt"] |

### 2.2.4 Stopword Removal

Stopword Removal is the step of removing unnecessary words from the input text [12]. Stopwords themselves have the same meaning as non-descriptive words in the text and can be removed. Table 5 is an illustration of a Stopword Removal sentence.

**Table 5.** Stopword Removal

| Tokenization results | Stopword Removal results |
|---|---|
| ["betty", "davis", "gives", "the", "finest", "performance", "of", "her", "career", "if", "there", "was", "ever", "any", "doubt", "in", "your", "mind", "that", "she", "was", "one", "of", "the", "greatest", "of", "screen", "actresses", "her", "performance", "here", "will", "dissipate", "ail", "doubt"] | ["betty", "davis", "gives", "finest", "performance", "career", "ever", "doubt", "mind", "one", "greatest", "screen", "actresses", "performance", "dissipate", "ail", "doubt"] |

### 2.2.5 Stemming

Stemming is the process of converting words contained in reviews into basic words according to certain rules [13]. Table 6 provides an illustration of Stemming.

**Table 6.** Stemming

| Stopword Removal results | Stemming results |
|---|---|
| ["betty", "davis", "gives", "finest", "performance", "career", "ever", "doubt", "mind", "one", "greatest", "screen", "actresses", "performance", "dissipate", "ail", "doubt"] | ["betti", "davi", "give", "finest", "perform", "career", "ever", "doubt", "mind", "one", "greatest", "screen", "actress", "perform", "dissip", "ail", "doubt"] |

Examples of the results of sentences that pass through the Preprocessing stage including Cleansing, Case Folding, Tokenization, Stopword Removal, and Stemming as shown in Table 7.

**Table 7.** Preprocessing Result

| Example sentences | Preprocessing results |
|---|---|
| "Betty Davis gives the finest performance of her career. If there was ever any doubt in your mind that she was one of the greatest of screen actresses, her performance here will dissipate ail doubt." | betti davi give finest perform career ever doubt mind one greatest screen actress perform dissip ail doubt |

### 2.3 Feature Extraction using TF-IDF

The research undertaken this time uses TF-IDF feature extraction to execute word weighting once the preprocessing stage is complete. TF-IDF feature extraction will convert words into numerical data. TF-IDF is a technique used in word processing to weight words in a document [14]. In this process, the inverse document frequency (IDF) and term frequency (TF) are calculated using equation from step 1.

**Building of Informatics, Technology and Science (BITS)**
Volume 5, No 1, June 2023 Page: 162−170
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i1.3595

$$W_{t,d} = tf_{t,d} \times idf_{t,d} \times \log\frac{N}{df_t} \quad (1)$$

Description:

| | |
|---|---|
| $w(t,d)$ | term(t) weight in document(d) |
| $tf(t,d)$ | number of times the phrase (t) appears in document (d) |
| $idf(t)$ | Inverse document frequencies for each word |
| $df(t)$ | amount of documents with each word's frequency |
| $N$ | number of papers overall |

### 2.4 Feature Selection using Information Gain

Information Gain approach is used for feature selection in this research following the feature extraction stage. Feature selection is the process of selecting attributes that are considered relevant or informative from a set of features in the data mining process. The number of attributes will affect the calculation, and even if many unrelated attributes are used in classification, it can also affect the results [15]. Information Gain has the advantages of being simple and easy to implement, can be assigned to numerical and categorical features. Information Gain is a fairly popular method for determining important criteria in text documents [5]. Information Gain can also help how relevant a feature is to the target label in the dataset. Information Gain feature selection will measure how much information about the presence and absence of a word plays a role in making the right classification decision in each class [13]. Using the equation in number 2, distracting or irrelevant features that will interfere with the categorization process are reduced.

$$\begin{aligned} IG(t) = & \sum_{i=1}^{|c|} P(c_i)\log P(c_i) \\ & + P(t)\sum_{i=1}^{|c|} P(c_i|t)\log P(c_i|t) \\ & + P(\bar{t})\sum_{i=1}^{|c|} P(c_i|\bar{t})\log P\,(c_i|\bar{t}) \end{aligned} \quad (2)$$

Description:

| | |
|---|---|
| $c_i$ | = categories |
| $P(c_i)$ | = odds of category |
| $P(t)$ | = possibility that term(t) will appear in document |
| $P(\bar{t})$ | = possibility that term(t) does not occur in document |
| $P(c_i\|t)$ | = conditional likelihood that a category will emerge given term(t) |
| $P(c_i\|\bar{t})$ | = the conditional likelihood that category at term(t) won't show up |

### 2.5 Split Data

Data were split into train data and test data for the investigation. A total of 80% of data is allocated for train data, while the remaining 20% is allocated for test data. Table 8 shows the outcomes of the data division.

**Table 8.** Total Train Data and Test Data

| Split Data | Data Train | Data Test |
|---|---|---|
| Total Data | 32589 | 8148 |
| Class Positive (Fresh) | 23209 | 5777 |
| Class Negative (Rotten) | 9380 | 2371 |

### 2.6 Logistic Regression Classsification

When the dependent variable is binary, Logistic Regression is appropriate in this study. Data can be described and connection between a binary dependent variable and one or more independent variables at a nominal, ordinal, interval, or ratio level explained using logistic regression [16]. The assumption made by logistic regression is that the independent features and the dependent variable's log-odds have a linear relationship. Logistic Regression is a classification method in statistical machine learning. In addition, it is also included in supervised learning methods [17]. The Logistic Regression method has Ordinary Least Square (OLS) techniques and procedures that are often used to estimate parameter values linearly [18]. Logistic Regression produces probabilities as output, which are useful in making decisions based on confidence levels. Logistic Regression generally contains yes/no, true/false values, or if in binary numbers, namely 1/0, it can be seen in the equation as in number 3.

$$Ln\left(\frac{\rho}{i-\rho}\right) = B_0 + B_1 X \quad (3)$$

Description:

**Building of Informatics, Technology and Science (BITS)**
Volume 5, No 1, June 2023 Page: 162−170
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i1.3595

$Ln$ = natural logarithm
$\rho$ = probability
$B_o + B_1X$ = an equation commonly known as Ordinary Least Square (OLS)

### 2.7 Evaluation

A system model's performance is measured by evaluation using a confusion matrix. To examine the factual data and prognostic outcomes of a categorization system, use a confusion matrix. Data matrices can typically be used to assess the effectiveness of a categorization system. Table 9 shows the confusion matrix table presentation.

**Table 9.** Confusion Matrix

| Confusion Matrix | | Factual Value | |
|---|---|---|---|
| | | **Positive** | **Negative** |
| **Value** | Positive | TP | FP |
| | Negative | FN | TN |

Description:
TP = positive predicted data and positive factual data (true positive)
FN = negative predicted data and positive factual data (false negative)
FP = positive predicted data and negative factual data (false positive)
TN = negative predicted data and negative factual data (true negative)

To assess how well the classification process is working, the confusion matrix calculates f1-score, precision and recall can be seen in equations number 4, number 5, and number 6.

F1-score is a harmonized average between precision and recall [19]. The formula of f1-score can be seen in equation number 4.

$$F1Score = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (4)$$

Precision is the percentage of number of positive class predictions to total number of positive and negative classes [20]. The formula of precission can be seen in equation number 5.

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

Recall is the percentage ratio of the actual class score for the positive class to the total of all positive and negative classes [20]. The formula of recall can be seen in equation number 6.

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

## 3. RESULT AND DISCUSSION

In the evaluation stage of the research conducted, it aims to test success of designed system to achieve final accuracy. The first stage carried out is preprocessing. Cleansing, Case Folding, Tokenization, Stopword Removal, and Stemming are all parts of the preprocessing. In this study, 40,737 datasets of English-language movie reviews that have been selected based on starting from 2020 and above are taken on the Rotten Tomatoes website which have passed the preprocessing stage or are clean, which then enter the TF-IDF extraction feature. After being weighted by the TF-IDF extraction feature, the data then enters the Information Gain feature selection stage with the number of features selected, namely k = 5000. Furthermore, with an 80:20 division ratio, the data will be split into train data and test data, with a total of 32589 train data, and 8148 for test data. This study's classification model, which employs Logistic Regression. The first test case is at the stage of evaluating preprocessing by examining the impact of utilizing stemming, to determine whether accuracy can be improved or decreased without stemming. The classification algorithm is not employing Information Gain feature selection in the second test case. This example demonstrates how feature selection can enhance this dataset's performance when used with the Logistic Regression technique. The final test scenario involves adding parameter values to the Logistic Regression modeling procedure. This scenario is to see how well it affects the final f1-score by adding parameters to the Logistic Regression algorithm. The experiments conducted for this investigation are displayed in Table 10 below.

**Table 10.** Experiment Scenario

| Experiment | Scenario |
|---|---|
| 1 | Utilizing and not using During the preprocessing stage, stemming |
| 2 | Using Information Gain and not using it to choose features |
| 3 | Adding parameter values to the Logistic Regression classification modeling process |

**Building of Informatics, Technology and Science (BITS)**
Volume 5, No 1, June 2023 Page: 162−170
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i1.3595

### 3.1 Scenario 1: Using and without using Stemming in preprocessing stage

In first experimental scenario, we used stemming and without using stemming in the preprocessing stage to find out how much it affects the final results of the model. Preprocessing was carried out twice in this experiment, once as shown in Figure 4 and again as shown in Figure 5.

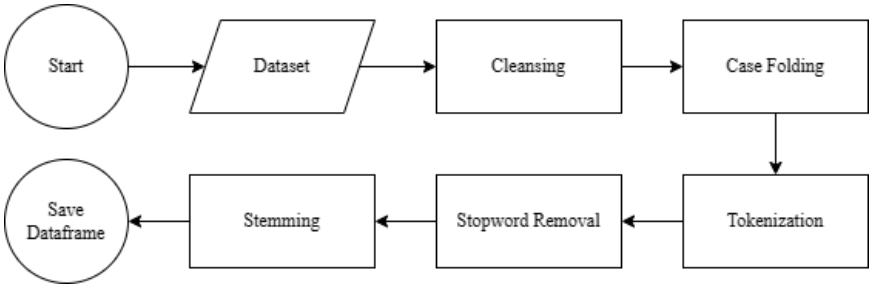


**Figure 4.** Preprocessing with Stemming

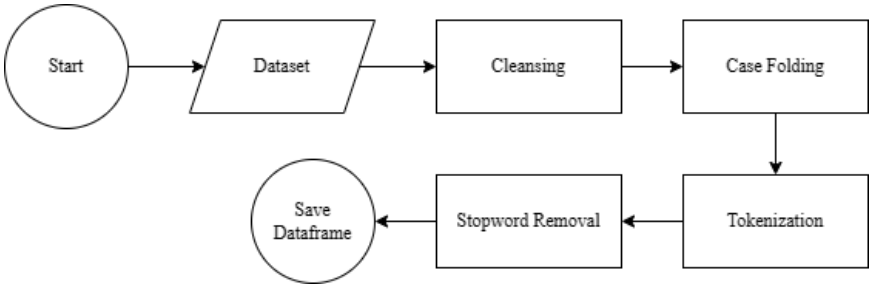


**Figure 5.** Preprocessing without Stemming

In the preprocessing stage, the two procedures differ by utilizing stemming and by not using stemming. "DataPreprocessingStemming" and "DataPreprocessingNoStemming" are the results of the data cleaning processes shown in the first and second figures, respectively. Both datasets are then weighted with TF-IDF then without using Information Gain as feature selection and modeled with Logistic Regression algorithm by adding parameter values. Table 11 presents the outcomes of this case.

**Table 11.** Effect without using Stemming at the Preprocessiong stage

| Preprocessing | Precission | Recall | F1-Score |
|---|---|---|---|
| With Stemming | 75.25% | 77.90% | 76.17% |
| Without Stemming | **75.56%** | **78.19%** | **76.50%** |

Based on results of scenario 1 testing in Table 11. it has shown that testing without using stemming in preprocessing stage produces slightly better f1-score, precision, and recall values when compared to using stemming in preprocessing stage. F1-score value obtained from the test results of scenario 1 is 76.50%, precision is 75.25%, and for recall is 78.19%. The reason for the decrease when using stemming in the preprocessing stage is change in meaning of the word which changes it to the base word. One example of a change in word meaning is "studies" and "studying" turns into "study", which can change the original word. This can cause loss of important information that causes model performance and a decrease in the f1-score value when compared to without using stemming which has slightly better performance results in the Logistic Regression method.

### 3.2 Scenario 2: Using and without using Information Gain for feature selection

It was examined without employing Information Gain feature selection in the second experimental situation. This experiment was carried out to determine how much feature selection for Information Gain impacts the model's output. This experiment uses an English movie review dataset, where the preprocessing process is done without using stemming. The data is analyzed using the Logistic Regression algorithm and adding the values of the parameters. Table 12 provides the experiment's findings.

**Table 12.** Effect without using Information Gain as a Selection Feature

| Selection Features | Precission | Recall | F1-Score |
|---|---|---|---|
| With Information Gain | 74.25% | 77.28% | 75.19% |
| Without Information Gain | **75.56%** | **78.19%** | **76.50%** |

Based on the results of testing scenario 2 in Table 12. has shown that testing without using Information Gain for its selection features produces slightly better f1-score, precision, and recall values when compared to using Information Gain as a selection feature. The f1-score value obtained from the scenario 2 test results is 76.50%,

**Building of Informatics, Technology and Science (BITS)**
Volume 5, No 1, June 2023 Page: 162−170
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i1.3595

precision is 75.25%, and for recall is 78.19%. The reason for the decrease when using Information Gain as a selection feature is that Information Gain focuses on reducing uncertainty or data diversity. However, there is a risk that relevant information for a particular classification task is lost as well. Some features that have important information or are correlated with the target may not be selected during feature selection with information gain, resulting in a deterioration of model performance and f1-score.

### 3.3 Scenario 3: Adding parameter values to the Logistic Regression classification modeling process

In scenario 3, it compares the effect of parameter content on Logistic Regression by not using default parameters (empty). This test is carried out by comparing contents of the parameters determined when conducting research, namely C = 1.2, penalty = 'l2', solver = 'sag', class_weight = 'balanced', random_state = 42 without using stemming at preprocessing stage and without using Information Gain as a selection feature. Results of scenario 3 testing can be seen in Table 8.

**Table 13.** Effect of adding values to Logistic Regression parameters

| Logistic Regression Parameter | Precission | Recall | F1-Score |
|---|---|---|---|
| (C=1.2, penalty='l2', solver='sag', class_weight='balanced', random_state=42) | 75.56% | **78.19%** | **76.50%** |
| Default parameter (empty) | **79.86%** | 69.72% | 72.02% |

Based on the results of testing scenario 3 in Table 13. has shown that testing by adding parameters that are determined and not default, namely C = 1.2, penalty = 'l2', solver = 'sag', class_weight = 'balanced', random_state = 42 can produce better f1-score and recall values when compared to the default parameters. The f1-score value obtained from the test results of scenario 3 is 76.50%, and for recall it is 78.19%. Meanwhile, better precision is obtained in the default Logistic Regression parameter. "C" is the inverse of regulation. The smaller the "C" value, the more the regulation will be set into the model built. A higher "C" value allows the model to better fit the train data, but beyond that it can also result in overfitting. This model uses C=1.2 from the default C=1.0. To reduce the chance of overfitting and help control the complexity of the model, a "penalty" is needed that matches the model built, in this model built using penalty = 'l2' and it is the default penalty of Logistic Regression which is good enough to handle this model case. solver='sag' this parameter can determine the solution algorithm used to find a solution, for the solver itself the default is 'lbfgs'. class_weight='balanced' is used to overcome class imbalance in the data which is very suitable for this model because after seeing the distribution of labels in the preprocessing process is unbalanced, the way out to get the best f1-score is to use class_weight='balanced' for the default parameter class_weight='none'. The random_state parameter is used to set the seed (initial random number) in order to obtain results that are always repeated, for the default parameter of random_state is 'none'.

## 4. CONCLUSION

Based on the outcomes of the experiments conducted for the Sentiment Analysis of Movie Reviews on the Rotten Tomatoes Site Using the Logistic Regression Method and Information gain Feature Selection by conducting three trials. The first experiment compared the effects of preprocessing stemming with and without it. The model was then ran without using Information Gain as a feature selection in the second trial. Additionally, in the third experiment, the Logistic Regression parameters were given values to assess how big of an impact adding parameters had on the model's final output. Conclusions obtained are as follows. Stemming process in preprocessing stage does not make performance increase in the model built. Testing in scenario 1 with data without stemming gives a slightly better f1-score when compared to using stemming. The use of Information Gain classification features with k = 5000 features selected has not been able to make the f1-score increase in this built model because Information Gain reduces data diversity. However, the selection of parameter values for Logistic Regression such as (C=1.2, penalty='l2', solver='sag', class_weight='balanced', random_state=42) is very suitable for the model built and can make the f1-score of 76.50%, recall of 78.19%, and precission of 75.56%.

## REFERENCES

[1] F. T. Laily and A. P. Purbantina, "Digitalisasi Industri Perfilman Korea Selatan Melalui Netflix Sebagai Alternatif Pasar Ekspor Film," Expo. J. Ilmu Komun., vol. 4, no. 2, p. 141, 2021, doi: 10.33021/exp.v4i2.1494.
[2] R. S. Sasmita, "Research & Learning in Primary Education Pemanfaatan Internet Sebagai Sumber Belajar," J. Pendidik. Dan Konseling, vol. 1, pp. 1–5, 2020.
[3] C. A. Putri, "Analisis Sentimen Review Film Berbahasa Inggris Dengan Pendekatan Bidirectional Encoder Representations from Transformers," JATISI (Jurnal Tek. Inform. dan Sist. Informasi), vol. 6, no. 2, pp. 181–193, 2020, doi: 10.35957/jatisi.v6i2.206.
[4] T. Chamidy, M. Informatika, U. Islam, N. Maulana, M. Ibrahim, and A. Mechanism, "Bidirectional GRU dengan Attention Mechanism pada Analisis Sentimen PLN Mobile," vol. 22, no. 2, pp. 358–372, 2023.
[5] R. I. Pristiyanti, M. A. Fauzi, and L. Muflikhah, "Sentiment Analysis Peringkasan Review Film Menggunakan Metode Information Gain dan K-Nearest Neighbor," vol. 2, no. 3, pp. 1179–1186, 2018.

**Building of Informatics, Technology and Science (BITS)**
Volume 5, No 1, June 2023 Page: 162−170
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i1.3595

[6] C. G. Kencana and Y. Sibaroni, "Klasifikasi Sentiment Analysis pada Review Buku Novel Berbahasa Inggris dengan Menggunakan Metode Support Vector Machine ( SVM )," vol. 6, no. 3, pp. 10451–10462, 2019.
[7] S. Priyanka and V. Ramya, "Classification Model To Determine the Polarity of Movie Review Using Logistic Regression," Int. Res. J. Comput. Sci. IRJCS Mendeley (Elsevier Indexed) CiteFactor J. Citations Impact Factor, vol. 1, no. 06, pp. 76–81, 2019.
[8] A. Syahadati, N. C. Lengkong, O. Safitri, S. Machsus, Y. R. Putra, and R. Nooraeni, "ANALISIS SENTIMEN PENERAPAN PSBB DI DKI JAKARTA DAN DAMPAKNYA TERHADAP PERGERAKAN IHSG," vol. 15, no. 1, pp. 20–25, 2021.
[9] B. Jonathan, J. I. Sihotang, and S. Martin, "Sentiment Analysis of Customer Reviews in Zomato Bangalore Restaurants Using Random Forest Classifier," vol. 7, no. 1, pp. 1719–1728, 2019.
[10] S. Wulan, U. Vitandy, A. A. Supianto, and F. A. Bachtiar, "Analisis Sentimen Evaluasi Kinerja Dosen menggunakan Term Frequency- Inverse Document Frequency dan Naïve Bayes Classifier," vol. 3, no. 6, 2019.
[11] A. Purnamawati, M. N. Winarto, and M. Mailasari, "Analisis Sentimen Aplikasi TikTok menggunakan Metode BM25 dan Improved K-NN Fitur Chi-Square," vol. 7, no. 1, pp. 97–105, 2023.
[12] A. Riyani, M. Zidny, and A. Burhanuddin, "Penerapan Cosine Similarity dan Pembobotan TF-IDF untuk Mendeteksi Kemiripan Dokumen," vol. 2, no. 1, pp. 23–27, 2019.
[13] A. B. P. Negara, H. Muhardi, and I. M. Putri, "Analisis Sentimen Maskapai Penerbangan Menggunakan Metode Naive Bayes dan Seleksi Fitur Information Gain," J. Teknol. Inf. dan Ilmu Komput., vol. 7, no. 3, p. 599, 2020, doi: 10.25126/jtiik.2020711947.
[14] R. Wati, S. Ernawati, and H. Rachmi, "Pembobotan TF-IDF Menggunakan Naïve Bayes Pada Sentimen Masyarakat Mengenai Isu Kenaikan BIPIH TF-IDF Weighting Using Naïve Bayes on Public Sentiment on The Issue of Rising BIPIH," vol. 13, no. April, pp. 84–93, 2023.
[15] Z. N. Syarif, "Penerapan Information Gain dan Algoritma K-Means Untuk Klasterisasi Kedisiplinan Pegawai Menggunakan Rapidminer," vol. 13, no. 1, pp. 1–12, 2023, doi: 10.36350/jbs.v13i1.165.
[16] M. Metode, K. N. Dan, and L. Regression, "Implementasi data mining untuk memprediksi penyakit jantung menggunakan metode k-nearest neighbor dan logistic regression," vol. 5, pp. 493–501, 2022, doi: 10.37600/tekinkom.v5i2.698.
[17] M. Shandy, T. Putra, and Y. Azhar, "Perbandingan Model Logistic Regression dan Artificial Neural Network pada Prediksi Pembatalan Hotel," vol. 6, no. 1, pp. 29–37, 2021.
[18] A. Novantika, "Analisis Sentimen Ulasan Pengguna Aplikasi Video Conference Google Meet menggunakan Metode SVM dan Logistic Regression," Prism. Pros. Semin. Nas. Mat., vol. 5, pp. 808–813, 2022.
[19] Y. S. HARIYANI, S. HADIYOSO, and T. S. SIADARI, "Deteksi Penyakit Covid-19 Berdasarkan Citra X-Ray Menggunakan Deep Residual Network," ELKOMIKA J. Tek. Energi Elektr. Tek. Telekomun. Tek. Elektron., vol. 8, no. 2, p. 443, 2020, doi: 10.26760/elkomika.v8i2.443.
[20] D. Chrisinta and J. E. Simarmata, "Analisis Sentimen Penilaian Masyarakat Terhadap Pejabat Publik Menggunakan Algoritma Naïve Bayes Classifier Sentiment Analysis of Society Assessment of Public Officials Using Naïve Bayes Classifier Algorithm," vol. 12, no. 148, 2023, doi: 10.34010/komputika.v12i1.9638.