CHI - SQUARE AND INFORMATION GAIN FEATURE SELECTION FOR HOTEL REVIEW SENTIMENT ANALYSIS USING SUPPORT VECTOR MACHINE

[1]Nathanael Karunia, [2]Yonathan Purbo Santosa
[1,2]Program Studi Teknik Informatika Fakultas Ilmu Komputer,
Universitas Katolik Soegijapranata
[2]yonathansantosa@unika.ac.id

ABSTRACT

In the current era, it has become a trend for people to order tickets online through online booking sites and applications, both in terms of transportation such as planes, vacations such as tours, and also lodging such as hotels. To get a good hotel, you need a review from people who have booked it. With the reviews written by visitors to the site or mobile application, they will then be analyzed so that an output can be produced that can be useful. One of the analytical models that can be done is sentiment analysis. The purpose of this study is to find the best method in analyzing sentiment based on the preprocessing of the data and hopefully it can produce knowledge in the form of sentiment analysis classification methods in order to determine a good method devoted to the data preprocessing section. The algorithm used to make this sentiment classification analysis is the Support Vector Machine using 3 feature selection methods, namely not using the selection feature, using the chi square selection feature, and using the information gain selection feature. The process consists of five steps in this study, which include several activities. namely data collection, preprocessing, feature extraction, feature selection, classification, and calculating accuracy. In the process of calculating accuracy, I used the Confusion Matrix method to find the best method of the three based on the accuracy results obtained. The results of the 3 uses of the feature selection method that were carried out were using the chi square feature selection method, the highest results were obtained, namely with an average accuracy of 86.68% which was followed by the use of the information gain selection feature which obtained an average accuracy of 85.78% and the last one was followed by the method not using the selection feature which got an average accuracy of 85.24%. From the results of the three methods, it can be concluded that the use of the chi square feature selection method in the case of sentiment analysis on hotel reviews is the best compared to the other two.

Keyword: feature selection, sentiment analysis, hotel review

INTRODUCTION

Background

In the current era, it has become a trend for people to order tickets online through online booking sites and applications, both in terms of transportation such as planes, vacations such as tours, and also lodging such as hotels. Website services and mobile applications for ordering tickets, such as Tiket.com, Traveloka, Blibli, and many more. This site and mobile application is equipped with features that really help the customer in determining which hotel to choose as a place to stay in an area. The feature in question is a review containing various comments from customers who have used hotel room reservations. Prospective buyers can get a more objective picture with these reviews, making it easier for prospective buyers to choose a hotel as a place to stay. With the reviews written by visitors to the site or mobile application, they will then be analyzed so that an output can be produced that can be useful. One of the analytical models that can be done is sentiment analysis. There are several classification methods used for sentiment analysis such as Support Vector Machine, Naive Bayes, Character Based N-gram model, and many more.

This project performs a sentiment classification analysis with hotel reviews and ratings. Sentiment analysis is a computerized technology that can help and analyze a sentence of someone's opinion that is textual [1]. The way to work from sentiment analysis is to understand and extract it like text mining to produce sentiment information [1]. Sentiment analysis connects all data, where previously unstructured becomes structured [1]. In the pre-processing stage of sentiment analysis, there is a feature selection which is useful for selecting features that have been obtained at the feature extraction stage to find features that are very influential in the case study [2]. This feature selection is what I was looking for with the hotel reviews and ratings dataset.

One of the problems of the previous sentiment classification is that there is no feature selection used in the pre-processing process [1]. In the research that has been done in classifying sentiment using the Support Vector Machine algorithm and tf-idf feature extraction [1]. Therefore, in this study, I analyzed sentiment with the same process as previous research by adding and comparing feature selection, namely chi-square and information gain on hotel reviews and ratings [1]. Feature selection is a selection process subset of terms in the training set and used in text classification [3]. Feature selection has 2 main goals, namely, making training data used for more classifiers efficient by reducing size vocabulary, and to improve accuracy classification by removing noise features [3]. Information gain is a symmetrical measure, ie the amount of information obtained by Y after observing X is equal to the amount of information obtained by X after observing Y [2]. Symmetry is the desired property to measure correlated features [2]. While Chi-square is one of the capable supervised feature selection removes many features without reduce the level of accuracy [3]. In Chi- square feature selection based on the theory statistics, two events of which are, occurrence of features and occurrence of category, which then each term value sorted from highest [3].

In this project, I use the Support Vector Machine algorithm to calculate accuracy and flexibility in determining positive and negative sentiments given. Then, in feature extraction I use TF-IDF and in feature selection, I use chi - square and information gain. In addition, I also compared the pre-processing process, namely at the feature selection stage using chi - square and information gain. I use this selection feature to compare the results of the accuracy values when not using the selection feature [1] and also when using the selection feature and compare 2 selection feature methods, namely chi - square and information gain whereas when using chi - square [3] and using information gain [2] and not using the selection feature [1], which one has the higher accuracy value. Given the previous findings, this research will try to answer the following questions,

1. Can feature selection with chi–square can improve the accuracy of this sentiment analysis?
2. Can feature selection with information gain can improve the accuracy of this sentiment analysis?
3. Between chi-square and information gain, which one is better for analyzing sentiment with a hotel review dataset based on the accuracy results in this case study?

The dataset that I use and analyze is 20492 data that I got from Kaggle TripAdvisor's Hotel Reviews. I converted the dataset into a CSV file to make it easier to analyze. The factors that I analyze are reviews and ratings. There are 2 variables in the dataset, namely rating and review. I did 5 tests for each algorithm. Of the 5 tests, I distinguish the training and testing data with the percentage of training data starting from 80%, 70%, 60%, 50%, and 40%. The purpose of this study is to develop existing research using the Support Vector Machine algorithm and TF-IDF feature extraction, on feature selection by using chi-square and information gain and to find the best feature selection method between the two based on the accuracy produced in this study.

LITERATURE STUDY

Nomleni [1] conducted research on the classification of sentiment analysis about public complaints against the government on social media Facebook and Twitter. He uses 1 algorithm in this study, named Support Vector Machine (SVM). In this study, preprocessing was first carried out in several stages. The first is cleaning the document from unnecessary characters. Followed by parsing, which separates the free text into parts called sentences. Then tokenizing, here the sentences are broken down into words and change the uppercase letters to lowercase. Followed by stopword removal which in this process removes irrelevant words. Then, continued stemming to find the basic word of a word. Next, the weighting process is carried out with tf-idf where words are given weights based on the occurrence of words in each sentence. Then the next step is to enter the classification process, where the processed data will be entered into the Support Vector Machine algorithm, and the accuracy, precision, recall, and f-measure values are calculated to find out how good this algorithm is in performing sentiment analysis. In this study, 5 trials of the combination of training data and testing data were carried out. From these 5 experiments, the results obtained an average accuracy of 82%. This shows the use of the SVM algorithm is suitable for analyzing sentiment.

Santoso et al. [4] classifying Twitter users' perceptions of covid-19 cases using the Logistic Regression algorithm. The dataset obtained is a public dataset for 2020 sourced from the Kaggle website with the name Coronavirus tweets NLP - Text Classification with 41157 data. The dataset has 6 attributes, namely username, screen name, location, tweet at, original tweet, and sentiment. The attributes used are only original tweets because they contain tweet data from Twitter users which is the basis for classifying text data related to comments about Covid-19. In this study, 5 target labels classified emotions, namely Extremely Positive, Positive, Neutral, Negative, and

Extremely Negative. then the target class is grouped into 3 targets to facilitate data analysis. Accuracy and F1 scores are used to determine how accurate the model to be built is. First, data preprocessing is done, namely importing data files, analyzing data, visualizing data, preparing data, and making Logistic Regression models. Then proceed with removing mentions and hashtags, remove stopwords, tokenize, and build models. In the building model, the dataset is divided into 2, namely 80% training data with 32925 data and 20% testing data with 8232 data. Then, the calculation is carried out with the Logistic Regression algorithm. The final result of this research is that the test using the L2 hyperparameter gets an accuracy of 77% and gets an F1-score value of 74%, while the test using the Hyperparameter None gets an accuracy of 74% and gets an F1-score value of 70%. Thus, testing with the L2 hyperparameter is a test that can produce a better logistic regression model in determining the classification related to comments on Twitter about the coronavirus or Covid-19 being in the positive, negative, or neutral category.

Ade R. R. [2] performs a classification to predict the accuracy of student graduation by sentiment analysis using the Naive Bayes algorithm and using information gain feature selection. The parameters used in this study were gender, study program, semester 1 credits to semester 6 credits, and achievement index semester one to semester 6 achievement index. In this study, 10 times testing and data training was carried out using 10-fold cross-validation test. In the first stage, a 10-fold cross-validation test was carried out to make 10 combinations of training data and testing data, then pre-processing of data and feature selection, namely information gain, was carried out to find correlated features and discard irrelevant features. Furthermore, classification is carried out with Naive Bayes, the first step is to do is to calculate probabilities. Followed by multiplying all the variables. The results of the classification of sentiment analysis using Naive Bayes algorithm gets an accuracy value of 83% and Naive Bayes who use the Selection Feature Information Gain gets an average accuracy of 88% proving that using feature selection can increase the accuracy value in a case study of sentiment analysis.

Lestandy et al. [5] conducted a sentiment analysis on the COVID-19 vaccine tweet using the Reccurent Neural Network and Naive Bayes algorithms. The dataset used in this study is 5000 tweets of the covid-19 vaccine from Kaggle with the distribution of 3800 positive sentiment tweets, 800 negative sentiment tweets and 400 neutral sentiment tweets. This study uses RNN and Naive Bayes by adding the TF-IDF technique which aims to give weight to the word relationship of a document. The dataset obtained is then pre-processed data to optimize data processing. There are 4 stages of pre-processing, including remove punctuation, case folding, stemming and tokenizing. After that, it is continued with the TF- IDF calculation to give weight to the relationship of a word. This is done by combining 2 concepts, namely, the frequency of occurrence of a word in the document and the reverse frequency of the document containing the word. This research was successfully carried out by comparing the performance of several RNN and Naive Bayes methods using the TF-IDF weighting technique. The RNN method (TF-IDF) shows the best accuracy result, which is 97.77% compared to Naive Bayes (TF-IDF) with an accuracy value of 80%.

Somantri and Dairoh [6] analyze the sentiment of the assessment of tourist destinations based on text mining using the Naive Bayes algorithm and the Decision Tree algorithm. The dataset used is as many as 120 text file data with the contents of a different number of words in each file obtained from comments on google maps. The first stage begins with pre- processing the data to get the appropriate data by going through the tokenization process, stopwords, and data weighting. After the data is pre-processed, it is continued by using the Naive Bayes algorithm and Decision Tree on the data. The best level of accuracy of the Naive Bayes model is 77.50% using k-Fold = 8 by stratified sampling. For the highest level of accuracy by liner sampling is 72.50% using k-fold = 10, and by shuffled sampling using k- fold = 8 the highest level of accuracy produced is 74.17%. The highest level of accuracy obtained by using the Decision tree using the criterion parameter Gain_ratio by shuffled sampling is 60.68% with fold=7. For the results obtained using the Information gain criterion, the highest was 55.83% and the Gini Index the highest was 59.83%. Based on the results obtained, the method using Naive Bayes has a better accuracy rate than the decision tree, which has an accuracy rate of 73.33%.

Wankhade et al. [7] perform sentiment analysis on food reviews using Logistic Regression. The dataset used is food review data as much as 568454 data with 10 parameters, namely id, product id, user id, profile name, helpfulness numerator, helpfulness denominator, score, time summary, and text. For the dataset, there are 77% positive reviews and 23% negative reviews. The dataset obtained is pre-processed first to remove the symbol and make it into a numerical value with several stages, namely, tokenizing, counting, and normalizing. After that, it is included in the Logistic Regression algorithm. From the results obtained, sentiment analysis on food reviews using Logistic Regression was successful because the score obtained was quite high, with an average of 0.89.

Munarwan [8] implements sentiment analysis to determine the level of popularity in tourist destinations using the Naive Bayes algorithm. The dataset used is in the form of comment data from social media Facebook and Instagram and there are 5 variables used namely comment count, Facebook likes count, Facebook was here count, facebook talking about, and Instagram visitors. Each of these components is analyzed using sentiment analysis to determine whether a comment or opinion is positive, negative, or neutral. The first step is to collect point of interest (POI) data from tourist destinations that will be ranked. In this study, 50 POIs were set for tourist destinations on the island of Bali along with their latitude and longitude coordinates. Then, proceed with text pre-processing to process the irregularities of a text to be more structured or in other words to prepare the text so that it can be changed to be more structured. There are four stages of text preprocessing which include case folding, tokenizing, filtering, and stemming. After that, the data that has been pre- processed is continued with the Naive Bayes classification process. The purpose of the classification process is to determine whether a sentence is included as a member of the positive opinion class or as a member of the negative opinion class which is determined based on the larger Bayesian probability calculation value. If the result of the Bayesian probability of the sentence for the positive opinion class is greater then the sentence is included in the positive opinion category and vice versa. Based on the results of the experiments carried out there are several data on the accuracy of the sentiment obtained. The first 100 phrases with an accuracy of 0.657 (65.7%), the second 5000 phrases with an accuracy of 0.8267 (82.67%). From the data above, it can be seen that the more phrases that are owned as the core of the algorithm, the more accurate the sentiment analysis presented.

Somantri and Apriliani [3] implements sentiment analysis to help in decision making for finding suitable diner when traveling. Using Chi-square and information gain, feature selection was used and compared to find which method is better. As a result, information gain are able to achieve 72.45% accuracy, which is 3.08% higher when compared to Chi-square that only achieved 69.36% accuracy when used on SVM to classify the sentiment of user generated comments in Tripadvisor platform.

Sari and Wibowo [9] analyzing customer sentiment of JD.ID online store using the Naive Bayes Classifier algorithm based on emotion icon conversion. The dataset used is 900 tweets, 300 positive tweets, 300 neutral tweets, and 300 negative tweets. Tweet data collection on Twitter using RStudio by entering the Twitter API settings and typing the JD.ID keyword. This study uses the Naive Bayes Classifier (NBC) method with TF-IDF weighting accompanied by the addition of an emotion icon conversion feature (emoticon) to determine the sentiment class that exists from tweets about the JD.ID store. This research starts from pre-processing data (Case Folding, Convert Emoticon, Cleansing, Text Transformation, Stopword Removal, Tokenizing, and Stemming), TF-IDF weighting to avoid imperfect data, data interference, and inconsistent data. Furthermore, an analysis of the tweet data is carried out to see the sentiment contained in an opinion. The results of this study are the Naive Bayes Classifier without TF-IDF weighting and conversion of emotion icons (convert emoticon) has an accuracy value of 96.44% while the Naive Bayes Classifier with TF-IDF weighting and conversion of emotion icons (convert emoticon) has an accuracy value of 98%, thus an increase in accuracy of 1.56%.

Afrizal et al. [10] analyzing sentiment towards Jakarta residents about the presence of mass rapid transit using the Naive Bayes algorithm. The data was obtained from social media, namely Twitter with the keyword "MRTJakarta" during the public MRT trial period, namely from 5 - 23 March 2019. The tweets taken were 1000 tweets. From the collected tweets, labeling is done with 5 annotators. Consists of 2 types of labels/classes, namely positive and negative. Each annotator labels 1000 tweets. So there are 5 labels for each tweet. Followed by data pre-processing which at this stage aims to delete and clean data in the form of words before the feature weighting process is carried out. After pre-processing the data, proceed with the classification process. The classification process is carried out on the training set with the aim of making a model, and then this model is used for testing. The first stage is to separate training data and testing data. 80% for training (800 tweets) and 20% for data testing (200 tweets). After separating, the classification process is carried out using Naive Bayes. In this research, the Naive Bayes algorithm can predict sentiment from tweets that have been collected regarding public interest in MRT Jakarta with an accuracy of 75%. This shows that the Naive Bayes algorithm is able to analyze the sentiment of tweets related to the presence of mass rapid transit in Jakarta.

In this research, I have a reference as the basis for this research flow process [1]. With the same process, I distinguish it in the feature selection section wherein previous studies there was no selection feature [1]. I use 2 selection features, namely chi-square and information gain and I also compare the results of the two feature which one is the best based on the accuracy value generated in this case study. Then, I also varied the training data and the test data that I did on the composition with 5 trials.

RESEARCH METHODOLOGY

This research has several stages. First, get a suitable dataset for this research. Second, I did the pre-processing of the data. Third, classify the clean data on the Support Vector Machine (SVM) algorithm. In this study, the workflow is as follows:

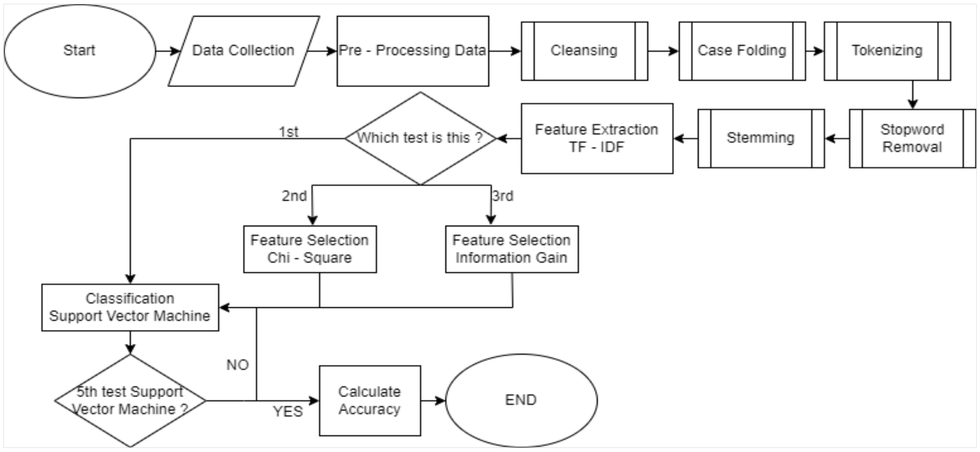
Start
Data Collection
Pre - Processing Data
Cleansing
Case Folding
Tokenizing
Stopword Removal
Stemming
Feature Extraction TF - IDF
Which test is this ?
1st
2nd
3rd
Feature Selection Chi - Square
Feature Selection Information Gain
Classification Support Vector Machine
5th test Support Vector Machine ?
NO
YES
Calculate Accuracy
END

Figure 1. Workflow

Data Collection

From the workflow on Figure 1, I started with the data collection process. The dataset that I used was taken from the Kaggle website in 2016. The dataset can be downloaded at the link https://www.kaggle.com/andrewmvd/trip-advisor-hotel-reviews has a file size 14.6MB csv file with 20,492 review data along with the rating figures that have been given by hotel users. This dataset has 3 attributes, namely, id, review, and rating. The example can bee seen in Table 1.

Table 1. Dataset

id	Review	Rating
1	Nice Beautifull Hotel Expensive parking got good deal stay hotel anniversary, arrived late evening took advice previous reviews did valet parking, check quick easy, little disappointed non-existent view room room clean nice size, bed comfortable woke stiff neck high pillows,….	4
2	ok nothing special charge diamond member hilton decided chain shot 20th anniversary seattle, start booked suite paid extra website description not, suite bedroom bathroom standard hotel room, took printed reservation…..	2

id	Review	Rating
3	nice rooms not 4* experience hotel monaco seattle good hotel n't 4* level.positives large bathroom mediterranean suite comfortable bed pillowsattentive housekeeping staffnegatives ac unit malfunctioned stay desk disorganized, missed 3 separate wakeup calls, concierge busy hard touch...	3
4	unique, great stay, wonderful time hotel monaco, location excellent short stroll main downtown shopping area, pet friendly room showed no signs animal hair smells, monaco suite sleeping area big striped curtains pulled closed….	5
5	great stay great stay, went seahawk game awesome, downfall view building did n't complain, room huge staff helpful, booked hotels website seahawk package….	5

Data Preprocessing

Preprocessing data is the first step in doing sentiment analysis. In this data preprocessing stage, the raw data is cleaned in several steps before being entered into the feature extraction stage and others. Preprocessing this data is divided into several stages as follows:

Document Cleaning Process (Cleansing)

At this stage, the data that I have found is cleaned. This cleaning is done to remove characters such as html, hashtags, website addresses, usernames (@), and punctuation marks (.,'":;[]!?%&()<>) which aims to reduce noise on data.

Case Folding

After the cleansing stage, the data is entered in the case folding stage where at this stage all the words contained in the review column will be converted into lowercase letters according to the letter and eliminate the characters in the review column because they can be considered as barriers. An example of the results of case folding can be seen below:

Tokenizing

At this stage, the data that has been carried out in the case folding stage will be continued with the tokenizing stage. In this tokenizing stage, each sentence in the review column will be broken down into words, then proceed with word collection analysis by separating the words and determining the syntactic structure of the data for each word.

Stopword Removal

After the tokenizing stage, the data is continued with the stopword removal process where at this stage the unimportant words are based on the stopword dictionary contained in the nltk.stopword library (English). These words are conjunctions such as (which, on, to, yes, no, etc.) or words that have no meaning will be deleted because they can affect the speed and performance of the classification later.

Stemming

At the stemming stage, the data that has been done with the stopword removal process will be continued at the stemming stage where at this stage the words in the review column are converted into basic words by removing affixes such as affixes, namely prefixes, insertions, suffixes, and combinations of prefixes and suffixes on derived words in the review sentence.

Feature Extraction

The next process after preprocessing the data is term-weighting. Term-weighting is the process of assigning term weights to the data in my case study, namely the data in the reviews column. The method I use to perform feature extraction in this research is TF-IDF (Term Frequency-Inverse Document Frequency).

TF (Term Frequency)

Term Frequency (TF) is a process to calculate the frequency of the number of occurrences of words in a dataset. Because the length of each sentence can be different, usually the value of TF will be divided by the length of the data (the sum of all data in the dataset).

$$tf_{t,d} = \frac{n_{t,d}}{Total\ number\ of\ terms\ in\ document} \tag{3}$$

Description :

tf = frequency of occurrence of words in a data

n = number of occurrences of words in the data

IDF (Inverse Document Frequency)

After we get the value of TF (Term Frequency), we continue to calculate from the value of IDF (Inverse Document Frequency). IDF is a counting process to be able to determine how important a word is in the dataset. IDF assesses words that often appear as less important words based on how they appear throughout the document. The smaller the value of this IDF, the less important the word is. Meanwhile, the greater the value of the IDF, the more important the word will be.

$$idf_d = \log\left(\frac{Number\ of\ Document}{Number\ of\ selected\ word\ frequency}\right) \tag{4}$$

TF-IDF (Term Frequency - Inverse Document Frequency)

After we get the TF (Term Frequency) and IDF (Inverse Document Frequency) values, we can calculate the TF-IDF value which is the product of the TF value and IF value. The TF-IDF formula can be seen below:

$$tfidf_{t.d} = tf_{t.d} \times idf_d \tag{5}$$

Feature Selection

The next process after the data feature extraction is performed is the feature selection process. Feature selection is the process of reducing irrelevant features and redundant data to select the best features from a feature data set. There are 2 methods that I use to perform feature selection in this research, namely Chi – Square and Information Gain. However, in this study I conducted 3 experiments at the feature selection stage, namely the first one did not use feature selection as in previous studies [1], used Chi-square, and used information gain.

Chi - Square

Chi-square is a feature selection method to test the relationship or effect of two variables and measure the strength of the relationship between one variable and another. Chi-square has a formula that can be seen below:

$$\chi^2 = \frac{\sum_i (O_i - E_i)^2}{E} \tag{6}$$

Description :

χ^2= Chi Square value

$O_i = F$ = Frequency of observed results (observed value)

E_i = Expected frequency (expected value)

Here I try some hyperparameters by taking 1000, 2000, 3000, 4000, and 5000. The results of this chi square value are directly transferred to the Support Vector Machine algorithm, however, I will explain the next feature selection first, namely Feature Selection using Information Gain.

Information Gain

Similar to chi-square, Information Gain is also a feature selection method that aims to determine attributes that will be used or discarded later. Information gain is carried out in several stages, namely calculating the information gain value for each attribute in the dataset, determining the desired threshold, and improving the dataset by reducing the attribute which is the purpose of this feature selection. Information gain has a formula that can be seen below:

$$Gain(A) = I(D) - I(A) \tag{7}$$

Description:

Gain (A) = Attribute Information A

I (D) = Total entropy

I (A) = Entropy A

Where to calculate the entropy (I (D) in the Information Gain formula above) the formula is obtained as follows:

$$info(A) = -\sum_{i=1}^{m} p_i \, log_2(p_i) \tag{8}$$

scription :

D = Case set

m = Number of partitions D

pi = Proportion of Di to D

And to calculate the entropy A (I (A) in the Information Gain formula above) has a formula like this:

$$info_A(D) = -\frac{\sum_{j=1}^{v}\left|D_j\right|}{|D|} \times I\left(D_j\right) \tag{9}$$

Description :

D = Case set

A = Attribute

v = Number of partitions A

|Dj| = Number of cases on j partition

|D| = Number of cases in D

I (Dj) = Total entropy in partition

At this stage, I tried several hyperparameters, namely with information gain values above -0.3, -0.2, and -0.1 to be used for the next classification stage. After performing the Information Gain feature selection stage, features with a high gain value will be obtained and become new features to be included in the algorithm using the Support Vector Machine as in previous research [1].

Classification

In this study, I use the Support Vector Machine (SVM) algorithm which is one of the methods in supervised learning which at this time I am using it for classification although it can also be used for regression. The data proportion for training and testing can be seen in Table 2.

Table 2. Data Composition

Data Compare	
Data Training	***Data Testing***
80%	20%
70%	30%
60%	40%
50%	50%
40%	60%

Support Vector Machine (SVM) classification method tries to find the best hyperplane function among an unlimited number of functions. Hyperplane is a function that can be used to separate between classes. In 2 dimensions, the function used in this hyperplane for classification between classes is called line whereas. The best hyperplane is the dividing line between 2 data classes in the input space which can be determined by measuring the margin of the hyperplane and finding its maximum point. Where margin means the distance between the hyperplane and the closest data from each class. The data closest to the hyperplane is called the support vector.

In Table 2, it is explained from the results of preprocessing that there are 20492 data divided into 3 sentiment classes, namely positive (2), neutral (1), and negative (0). The data that has been normalized before being entered into the Support Vector Machine (SVM) algorithm, the data is divided into 2, namely training data and test data performed during classification. In this test, the data is divided 5 times with different input training data and testing data.

The data that was trained and tested has been divided and each experiment is classified using the Support Vector Machine (SVM) algorithm with a kernel function that maps linear data so that it gets a new dataset of learning models in each experiment. The results of the learning model are classified by testing 5 times were in each test using a matrix with a size of 3 x 3 as a representative of the actual class and the predicted class.

The concept of the Support Vector Machine algorithm can be described as an attempt to find the best hyperplane that serves as a dividing line between the two classes. The best dividing hyperplane between the two classes is determined by measuring the hyperplane margin and finding its maximum point. Margin is the distance between the hyperplane and the nearest point or data in each class. The closest pattern is called the support vector. It is called the best hyperplane because it is located right in the middle between the two classes, while the plus sign is orange and the blue circle inside the black circle is called the support vector. Efforts to find this hyperplane is the most important part of the classification of the Support Vector Machine algorithm. To get the perfect hyperplane location, it can be defined by the following formula:

$$f(x) = w^T x + b \tag{10}$$

$$\begin{gathered}[(w^T . x_i)] + b \ \geq 1 \ for y_i = +1 \\ [(w^T . x_i)] + b \ \leq -1 \ for y_i = -1\end{gathered} \tag{11}$$

With the description of xi as the training data set, i as 1,2,…., n , and yi as the class label of xi . The largest margin can be found by maximizing the value of the distance between the hyperplane and its closest point and can be formulated as a quadratic programming problem which means finding the minimum point using Lagrange multiplier.

$$\min_{\vec{w}} \tau(w) = \frac{1}{2} \|\vec{w}\|^2 \tag{12}$$

$$L(\vec{w}, b, a) = \frac{1}{2} \|\vec{w}\|^2 - \sum_i \alpha_i (y_i((\vec{w}, \vec{x}_i + b) - 1) \tag{13}$$

$$\sum_{i} \alpha_i - \frac{1}{2} \sum_{i,j=1} \alpha_i y_j \alpha_J y_i \vec{x}_i \vec{x}_j \tag{14}$$

In equation (12)-(14), α_i is the Langrange Multiplier which can be 0 (zero) or positive $\alpha_i \geq$ 0 (zero). The optimal value of the formula below can be seen by looking at the value of L against $\vec{w}$ and b and can maximize the value L against α_i. Based on basically the optimal point L = 0, the formula contained in equation (13) can set the maximization of the problem which only contains αi.

From the results of the Langrange Multiplier on above calculations can be obtained the value of αi which is positive. Data that is related or correlated with α_i positive is what can be called a support vector which is on Support Vector Machine algorithm.

$$\left(\alpha_i \geq 0 (i = 1,2, \dots, l) \quad \sum_{i=1}^{l} \alpha_i y_i = 0 \right) \tag{15}$$

To measure the classification performance on the original data and the result data from the classification model that has been done, you can use the confusion matrix which we will discuss in the next stage.

Confusion Matrix

Confusion matrix is a process of measuring the performance / performance of the classification model where the output can be in the form of 2 or more classes. There are four terms that can be said to be representative of the results of the classification process in the confusion matrix, namely True Positive, False Positive, True Negative and False Negative. From the results of these 4 categories, the values of accuracy, precision, recall, and F-1 Score can be calculated. Accuracy is how accurate the model is in classifying correctly. Precision is the accuracy between the requested data and the prediction results provided by the model. Then, recall is the success of the model in rediscovering information. While the F-1 Score is the average comparison between precision and recall which is weighted. These four things can be formulated as follows:

$$\text{Accuracy} = \frac{\text{TP}}{\text{Number of data}} \tag{18}$$

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{19}$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{20}$$

$$\text{F1 Score} = \frac{(\text{Precision} + \text{Recall})}{2} \tag{21}$$

At the feature extraction stage, it is carried out 3 times with a change of method. The first one does not use feature extraction, the second uses chi-square feature extraction, and the third uses information gain feature extraction. Each of these three methods was carried out 5 times with the distribution of different datasets as given above. After doing everything, I can determine which method is the best of all.

RESULTS

From the process that we have done above using the Support Vector Machine algorithm and the Information Gain selection feature as well as the Chi Square selection feature with the aim of increasing the accuracy of previous research [1].

Table 3. Compare Accuracy of All Methods Results

Test	No Feature Selection	Chi - Square	Information Gain
1	85.6%	87.1%	86.2%
2	85.5%	87.0%	86.0%
3	85.3%	86.7%	85.9%
4	85.0%	86.5%	85.5%
5	84.8%	86.1%	85.3%
MEAN	85.24%	86.68%	85.78%

Table 4. Compare F1-Score of All Methods Results

Test	**F1-Score**		
	No Feature Selection	***Chi - Square***	***Information Gain***
1	0.66	0.68	0.67
2	0.67	0.68	0.68
3	0.65	0.67	0.66
4	0.64	0.67	0.66
5	0.64	0.66	0.65

From the comparison results above, it can be seen that the sentiment analysis classification process using the Support Vector Machine algorithm without using the selection feature has the lowest level of accuracy compared to using the chi-square and information gain selection features. Then, the F1-score value on the neutral label in all methods got a low value, this is because the data after being labeled got little data. Because the formula for precision is the true positive value divided by the true positive value plus the false positive, where the value of true positive on the neutral label gets a low value so the precision value of the neutral label is low. And in this calculation the chi-square feature selection method is the best method in the sentiment analysis case study that I did with the highest accuracy compared to the other two. After getting the three comparisons, namely not using the selection feature, chi-square, and information gain, it can be concluded that the chi-square has the highest value and is followed by information gain and does not use the selection feature in this study and proves that using the selection feature can add value to the accuracy of previous studies [1].

CONCLUSION

From the classification process for sentiment analysis using the Support Vector Machine algorithm and the chi – square selection feature and information gain on hotel reviews to increase the accuracy of previous research [1] that I have done, it can be said that:

From classification calculations to sentiment analysis with the Support Vector Machine algorithm and using Chi – Square feature selection, it can increase higher accuracy than using the Support Vector Machine algorithm without using feature selection as in previous studies [1]. Where when using a method like the previous research [1] without using feature selection I get an average accuracy of 85.24% while when I use the chi - square feature selection I get an average accuracy of 86.68%. The use of chi-square feature selection resulted in an average accuracy increase of 1.44%. This proves that the use of chi-square feature selection can increase the accuracy of this sentiment analysis. So for case studies on both methods between using the chi-square selection feature and without using the selection feature, the chi-square can increase the accuracy of the previous research method [1].

Classification calculations for sentiment analysis with the Support Vector Machine algorithm and Information Gain feature selection can increase accuracy higher than using the Support Vector Machine algorithm without using feature selection as in previous studies [1]. Whereas when not using feature selection as in previous research [1], the average accuracy is 85.24%. When I use feature selection using the Information Gain method, it can increase the average accuracy by 85.78%. The use of feature selection with Information Gain can increase the average accuracy by 0.54%. This proves that using the Information Gain feature selection can increase the accuracy of the previous research method [1] in this sentiment analysis. So for a case study between these two methods, namely with the information gain selection feature and without using the selection feature, information gain can increase the accuracy of the previous research method [1].

In the classification results for sentiment analysis in hotel reviews using the Support Vector Machine algorithm, it can be seen that between using the chi-square selection feature and information gain, the accuracy value of each experiment proves that the use of the chi-square selection feature for classification in this hotel review is better. higher than the use of the information gain selection feature. Whereby using the chi-square selection feature, the average accuracy is 86.68%, while using the information gain selection feature, the average accuracy is 85.78%. From the two results, the average accuracy value has a distance difference of 0.90%. This proves that using the chi-square selection feature gets higher results than using information gain. The best method between information gain and chi-square in the classification in this study is chi-square. So it can be concluded that the use of the chi-square selection feature is better than the use of information gain in this research classification.

Further research can look for datasets with more attributes to see the results of the classification which may increase the accuracy of each method. Also, added a variety of feature selection methods to compare the process and results of the calculation of accuracy in the classification. Further research can use other classification algorithms such as Naive Bayes, Logistic Regression, and others to see the results of the classification which may increase the accuracy of each algorithm.

REFERENCES

[1] P. Nomleni, “SENTIMENT ANALYSIS MENGGUNAKAN SUPPORT VECTOR MACHINE(SVM),” Thesis, Institut Teknologi Sepuluh Nopember, Surabaya, 2015. Accessed: Aug. 01, 2023. [Online]. Available: https://repository.its.ac.id/41821/1/2213206717-Master%20Thesis.pdf

[2] A. R. Rozzaqi, “Naive Bayes dan Filtering Feature Selection Information Gain untuk Prediksi Ketepatan Kelulusan Mahasiswa,” *Jurnal Informatika Upgris*, vol. 1, no. 1 Juni, Art. no. 1 Juni, 2015, doi: 10.26877/jiu.v1i1.

[3] O. Somantri and D. Apriliani, “Support Vector Machine Berbasis Feature Selection Untuk Sentiment Analysis Kepuasan Pelanggan Terhadap Pelayanan Warung dan Restoran Kuliner Kota Tegal,” *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 5, no. 5, Art. no. 5, Oct. 2018, doi: 10.25126/jtiik.201855867.

[4] A. K. Santoso, A. Noviriandini, A. Kurniasih, B. D. Wicaksono, and A. Nuryanto, “KLASIFIKASI PERSEPSI PENGGUNA TWITTER TERHADAP KASUS COVID-19 MENGGUNAKAN METODE LOGISTIC REGRESSION,” *Jurnal Informatika Kaputama (JIK)*, vol. 5, no. 2, Art. no. 2, Jul. 2021.

[5] Merinda Lestandy, Abdurrahim Abdurrahim, and Lailis Syafa’ah, “Analisis Sentimen Tweet Vaksin COVID-19 Menggunakan Recurrent Neural Network dan Naïve Bayes,” *RESTI*, vol. 5, no. 4, pp. 802–808, Aug. 2021, doi: 10.29207/resti.v5i4.3308.

[6] O. Somantri and D. Dairoh, “Analisis Sentimen Penilaian Tempat Tujuan Wisata Kota Tegal Berbasis Text Mining,” *JEPIN*, vol. 5, no. 2, p. 191, Aug. 2019, doi: 10.26418/jp.v5i2.32661.

[7] M. Wankhade, A. C. S. Rao, S. Dara, and B. Kaushik, “A Sentiment Analysis of Food Review using Logistic Regression,” *int. j. sci. res. comput. sci. eng. inf. technol.*, vol. 2, no. 7, pp. 251–260, Sep. 2017, doi: 10.32628/CSEIT174430.

[8] M. Murnawan, “PEMANFAATAN ANALISIS SENTIMEN UNTUK PEMERINGKATAN POPULARITAS TUJUAN WISATA,” *Jurnal Penelitian Pos dan Informatika*, vol. 7, no. 2, p. 109, 2017.

[9] F. V. Sari and A. Wibowo, “ANALISIS SENTIMEN PELANGGAN TOKO ONLINE JD.ID MENGGUNAKAN METODE NAÏVE BAYES CLASSIFIER BERBASIS KONVERSI IKON EMOSI,” *Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer*, vol. 10, no. 2, Art. no. 2, Nov. 2019, doi: 10.24176/simet.v10i2.3487.

[10] S. Afrizal, H. N. Irmanda, N. Falih, and I. N. Isnainiyah, “Implementasi Metode Naïve Bayes untuk Analisis Sentimen Warga Jakarta Terhadap,” *Informatik : Jurnal Ilmu Komputer*, vol. 15, no. 3, p. 157, Aug. 2020, doi: 10.52958/iftk.v15i3.1454.