CHAPTER 3
RESEARCH METHODOLOGY

3.1. Data Collection

The data source that I use is taken from the Kaggle website, namely Trip Advisor Hotels Reviews in 2016. The dataset is a CSV file where there are 20,942 data in the form of reviews and ratings. This dataset will be divided into training data and testing data.

3.2. Preprocessing Data

The first thing to do is to preprocess the data because this stage is very important which aims to clean the data into the desired form. Therefore, at this stage there are several stages:

3.2.1. Document Cleaning Process (Cleansing)

Review sentences cleaned of characters such as html, hashtags, website addresses, username (@), punctuation marks (.,'":;[]!?%&()<>) and characters other than the alphabet to reduce noise.

3.2.2. Case Folding

At this stage, sentence reviews are changed from capital letters to lowercase letters and the letters are uniform from A to Z and other than letters will be removed because they are considered delimiters.

3.2.3. Tokenizing

In this tokenizing stage, the review sentence will be broken down into words and then analyze the collection of words by separating the words and determining the syntactic structure of the data for each word.

3.2.4. Filtering / Stopword Removal

In this stopword removal stage, words that are not important based on the stopword dictionary such as connecting words (which, on, to, yes) or words that have no meaning are removed.

3.2.5. Stemming

This stage converts words into basic words by removing affixes, namely prefixes, insertions, suffixes, and combinations of prefixes and suffixes on derived words in the review sentence.

3.3. Feature Extraction

This stage is a process for calculating and providing information on TF (Term-Frequency), DF (Document Frequency), and IDF (Inverse Document Frequency) by calculating review sentences based on the frequency of occurrence of terms in the dataset. The term calculates the probability of occurrence in the dataset.

3.4. Feature Selection

Feature selection is a technique to select important and relevant features to the data and reduce irrelevant features. Feature selection aims to select the best feature from a feature data set. At this stage I did 3 processes, namely not using feature selection as in previous studies [1], using chi-square, and using information gain.

3.4.1. Chi - Square

Chi-Square is a feature selection method which is included in the filter method. The following is the formula used to apply the method feature selection chi - square :

$$x^2 = \sum \frac{(O - E)^2}{E} \quad (1)$$

Description :

X^2 = chi squared

O = observed value

E = expected value

3.4.2. Information Gain

This information gain can reduce feature dimensions by measuring the entropy reduction before and after separation. The form of the formula for this information gain is as follows :

$$IG(S,A) = Entropy(S) - \sum_{c \in values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (2)$$

Description :

S = Case Set

A = Attribute

$|S_v|$ = Number of Cases on Partition v

$|S|$ = Number of Cases in S

3.5. Classification

Support Vector Machine can perform classification of more than two classes. For example in determining the sentiment of a reviews. There are three sentiments, namely positive, negative, and neutral. The first step of an SVM algorithm is to define the equation of a separating hyperplane. Hyperplane is a function that is used as a separator between one class and another. This function is used to classify inside a higher dimensional class space. In 2-dimensional form, the function used to classify between classes is called a whereas line. Once determined, the distance between the line and the support vector will be calculated. This distance is called the margin. The purpose of this algorithm is to maximize the existing margins, so as to get the optimal line/hyperplane.

3.6. Analysis

In this study, I tested the dataset 5 times with a combination of a large percentage of training data and testing data.

Table 3.1. Analysis

No.	Data Training	Data Testing
1	80%	20%
2	70%	30%
3	60%	40%
4	50%	50%
5	40%	60%

In those 5 experiments, I did 3 analyzes, the first one did not use feature selection [1], the second one used chi-square feature selection [8], and the third one used information gain

feature selection [3]. Of the three methods, I looked for the values of accuracy, precision, recall, and f- measure.