CHAPTER 1
INTRODUCTION

1.1. 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, Charcter 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 [3]. 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 [8]. 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 [8]. 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 [3]. Symmetry is the desired property to measure correlated features [3]. While Chi-square is one of the capable supervised feature selection removes many features without reduce the level of accuracy [8]. 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 [8].

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 [8] and using information gain [3] and not using the selection feature [1], which one has the higher accuracy value.

1.2. Problem Formulation

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 ?

1.3. Scope

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%.

1.4. Objective

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.