CHAPTER 1
INTRODUCTION

1.1. Background

Sentiment analysis is a grouping of text data generated from users or the process of analyzing the opinions of online users to determine the emotions contained in their writing. The internet is always filled with data or information in the form of text from various sources, such as films in the form of reviews. It is this analytical sentiment that plays a role in connecting all of these data, which were previously unstructured to become more structured. Sentiment movie analysis uses reviews to analyze user reviews. Reviews from users become a place to rate how good the film is, in the form of criticism or suggestions or praise.

In the analysis of the movie classification of a film, users tend to see whether a film is good or not from the film's rating. Most users are not interested in seeing other user reviews before, even though the rating is always tied to the reviews given by users to find out whether the film is good or not. Due to the large number of reviews left by users on a film website, users have difficulty knowing the classification of positive reviews and negative reviews, and users also do not know the quality of a film.

Therefore, in this project, a classification analysis or sentiment analysis is carried out for the film against a relevant review or rating, which is classified based on polarity. From the polarity, it can be seen which reviews tend to be positive or which tend to be negative. It is impossible to do a review analysis for a film manually, it requires a computer approach using machine learning. The large number of reviews that exist on a film needs a sorting process. The process to sort out positive reviews and negative reviews requires an algorithm. The algorithms that will be implemented in this project are Random Forest Classifier and Logistic Regression.

In this project, I made a sentiment analysis for a movie review. By using 2 algorithms, namely Random Forest Classifier and Logistic Regression to have a high value of accuracy and flexibility. Both algorithms determine positive and negative sentiments. The data structure used is a database, which will train data and test data for comparison purposes. The dataset is also converted into CSV to make it easier to analyze. I also compare which algorithm is better based on its accuracy.

1.2. Problem Formulation

1. Does the implementation of the Random Forest and Logistic Regression algorithms on sentiment analysis in movie reviews tend to be positive or negative?
2. How do the results compare the accuracy of the two algorithms? Which algorithm has the best accuracy?

1.3. Scope

The dataset used and analyzed is 2.000 data obtained from Kaggle IMDb Movie Reviews released in 2021, in the form of an SQLite database. The dataset is also converted into CSV to make it easier to analyze. The factors that I will analyze are only based on the variables, ratings, and reviews. There are 5 variables in these datasets namely id, review, rating, author, and title.

1.4. Objective

The purpose of this project is to analyze reviews into negative sentiments or positive sentiments by implementing the Random Forest and Logistic Regression algorithms for Sentiment Analysis on Movie Reviews and seeing the comparison between the two algorithms based on its accuracy.