# CHAPTER 3
# RESEARCH METHODOLOGY

## 3.1 Data Collection

In this project, the dataset that used is IMDb Movie Reviews 2021 which can be taken from Kaggle. This dataset has a data structure in the form of an SQLlite database with 2.000 data. This dataset contains reviews from users on movies released in 2021. There are 5 variables in this dataset, namely id, review, rating, author, and title. The dataset can be obtained by downloading a 4.17 MB zip file.

This dataset is broken down into training and test sets for comparison. The training and test datasets both have positive and negative reviews. The dataset has user reviews, each review has a rating from users on a scale of 1 to 10. Here the author converts the SQLlite dataset to CSV and changes the rating scale to a binary label. If the movie rating is more than 5 then the review will be said to be positive, if it is less than 5 it will be said to be negative.

## 3.2 Algorithms

The existence of a lot of data must be classified for positive reviews and negative reviews. This project will use the Random Forest and Logistic Regression algorithms. The author uses these 2 algorithms because they do not require hyperparameters and can have high accuracy and flexibility, and can analyze and determine which reviews are positive and which reviews are negative. Both algorithms are suitable for review classification so that training and test data becomes easier. Random Forest Classifier is a popular algorithm that is very suitable to be applied in the classification of review analysis.

## 3.3 Design

Data is cleaned and preprocessed due to raw text. For preprocessing, reviews are labeled based on rating. TF-IDF (Term Frequency-Inverse Document Frequency) is performed to count every word in the document. Then split the data which produces Training Data, Training Label, Testing Data, and Testing Label for the calculation of Count Training and Count Testing. Preprocessing should be classified into positive reviews and negative reviews. All elements that have breaks, stopwords are removed from the data because they do not show information about the impressions of user reviews on the film. Punctuation marks are not removed because it is possible to know the expression of the review. Then stemming is done which removes words ending in morphology. Vectorization technique is applied to convert text into matrix features, so the data can be understood by Machine Learning algorithms. This project implements 2 algorithms for the train set and test set of the feature matrix, namely Random Forest and Logistic Regression. 70% of the data was done by training and 30% by doing a test set. The test data will be used in the process of calculating the accuracy of the 2 algorithms after studying the train set.

## 3.4 Coding

This project uses Python 3.7 to perform the computations. The author uses Python because it can be learned easily and has automatic memory management. Python also has many tools that are suitable for machine learning. Python provides many complete libraries and frameworks that can be downloaded for free.

Python is also proven to be the best programming language for Machine Learning. In addition, Python also provides visualization results in its programming so that users can easily analyze data and code.

## 3.5 Analysis

This project uses 2.000 user review data from various movie titles. Of the 2.000 reviews, an analysis will be carried out to determine the data including positive sentiment or negative sentiment. Selection of the right algorithm is needed for the classification analysis process from user reviews. Which algorithm is the most precise, efficient, and has the best and most accurate accuracy. Moreover, Random Forest and Logistic Regression do not require hyperparameters and are suitable to be applied to classification. For performance training, accuracy is calculated for each algorithm.