IMDB Dataset of 50k Movie Reviews

# PROJECT PROPOSAL:

**Introduction & Background:**

Social media has significantly grown in popularity over the previous few decades. Everybody utilises social media in their daily lives and posts their feelings and ideas about various topics. Many businesses use the data for processing and produce relevant information from them that may be used in the future for commercial objectives. On websites like IMDB, Amazon, Rotten Tomatoes, and others, there is a vast amount of text-based data on movie reviews provided by the public, and manually analysing that data is a difficult operation. Sentimental analysis is utilised to resolve this issue and make the work simple and automated.The Sentimental analysis is a subfield of Natural Language Processing and AI, used for extracting meaningful information from the textual data which we can utilise further to improve business in all way.

In this project, I have taken dataset of IMDB movie reviews to predict how the audience have rated the movies and predict the movies that have a positive or negative review. I will create a model which I will used for extracting information using Sentiment Analysis from the data and predict which classifier is best suitable for this dataset by looking at the accuracy. The dataset is collected from Kaggle and it has 50k movie reviews.

# About Dataset:

IMDB Dataset consists of 50k movie reviews for natural language processing or Text analytics. The dataset has two columns Review and Sentiment. The Review filed stores the textual feedback given by the audience and Sentiment contains the pre-labeled value with “Positive” or “Negative”. The whole dataset is equally divided into 25000 for training and 25000 for testing set.

# Problem Statement:

The main goal is to analyze the sentiment of movie reviews and based on that I will classify the movie review as “Positive” or “Negative” and identifies best classifier for this dataset by looking at the accuracy.

Miren Patel

1

# Proposed Framework:

In this project, I have used Sentiment Analyses technique. The proposed framework for models includes following steps:

* Data Cleaning
* Text Stemming
* Removing Stop words and special characters:
* Feature Extraction
* Modelling
* Accuracy of the model
* Visualizing the classification report

**Accomplishment:**

The main purpose behind this project is to create a sentiment analysis model, which will help for better understanding of movie reviews and compares the various classifier’s result based on the accuracy observed from each model.

# References:

**https://**[**www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews?**](http://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews) **select=IMDB+Dataset.csv**

**https://chaitanya1731.github.io/img/prj-1/report.pdf**

Miren Patel

2