Exploring the Impact of Movie Reviews on Overall Success

Tanjim Reza, Fahad Al Mannan, Nafiz Siddiqui Adnan,
Md. Mustakin Alam, Md Sabbir Hossain, and Annajiat Alim Rasel
Department of Computer Science and Engineering (CSE)
Brac University

***Abstract*—Movie reviews tell us if a movie is loved by the audience or not which determines the box office revenue of the movie. The movie review and ratings contain the sentiments of the public and a numerical rating to share their opinion. In this study, We will analyze the connection between movie reviews on success, and different machine learning methods are used to determine the effectiveness of reviews on box office collections. Our model is tested using a real-world dataset and the methods to analyze and compare the models. The outcome of the research offers important implications for the film industry and sets the stage for future research in this area.**

I. Introduction

Movie reviews might play a vital role in determining the total revenue/income of a movie. We can learn a lot by analyzing the movie reviews and the movie revenue together which can be used to make better movies that are more appealing to the audiences. The main goal of this research is to analyze the reviews of the audience on IMDB and Metascore and classify them into different categories. These data will be processed and used with different machine-learning techniques to extract the outcomes.

The purpose of this research paper is to analyze the reviews of the audience on IMDB and Metascore and classify them into different categories. The review data may be used to determine the audience's opinion of the movie. Analyzing the perceptions will help us determine the kinds of movies people prefer. Moreover, we are using IMDB comments and Metascore reviews because IMDB has a vast source of opinions on newly released movies as well as old ones. The audience's review includes a wide variety of perspectives including personal opinions, political views, one's mental state, and box office collection. Metascore additionally includes critic reviews which have a massive impact on the audience. Our study focuses on these reviews and box office revenue. Machine learning techniques are employed to extract outcomes from the data.

- Extract relevant data from the reviews
- Train the machine
- Generate scores using classification algorithms

II. Literature Review

We have reviewed many movie reviews and revenue-based works before working on this paper. Among them, Li-Chen Cheng, a professor of Information and Finance Management from the National Taipei University of Technology in Taiwan showed remarkable work. She utilized Baseline and Extended Regression Models in her research. Her study on the effect of movie reviews on box office revenue deduced that the numerical ratings do not affect the revenues much, but positive and negative reviews had a significant impact. Another noticeable work we came across was Abdul Meral's analysis, who is a data scientist at LC WAIKIKI in Istanbul, Turkiye. He utilized deep learning and a neural network model to analyze movie reviews, where he got 89.99Liang et al. (2015), and Hu et al. (2018) used sentiment analysis and aspect-based sentiment analysis methods with machine learning techniques to determine the polarity of a sentence whether it expresses positive, negative, or neutral sentiment.

III. Dataset

The dataset we are working with has around 5000 movie data with IMDB rating scores, the number of critics reviewed, the number of user reviews, and box office revenue, budget, genre, etc. Movies with an IMDB rating over 7 are considered Positive, and a rating less than 5 is considered Negative. The numbers in between are labeled as Neutral. We have analyzed the movies that have at least 1000 voters.

IV. Methodology

Here, we have decided to use the Decision tree, Random Forest and Linear Regression to train our data. We will implement these machine learning algorithms to train and test our models respectively. Moreover, we are organizing the dataset we are using. We have done dataset balancing, preprocessing (null checking, stemming, etc.), and dataset splitting. Next, we plan to train the ML models using our dataset.

The steps we are following are illustrated in the "Fig: 1".

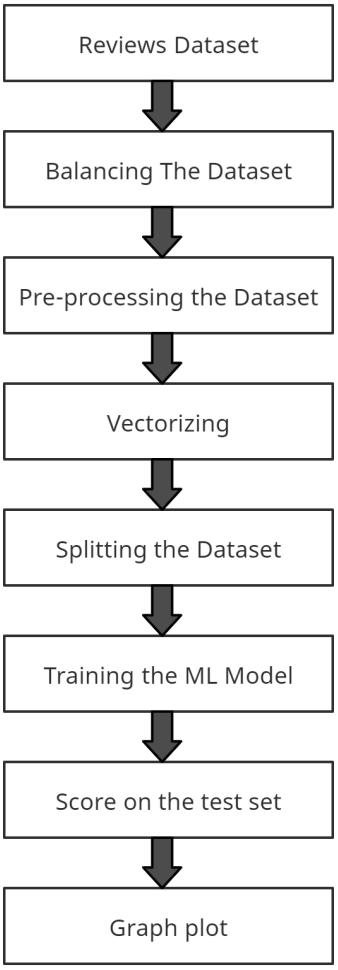

Fig. 1. Steps Followed

A. Balancing the Dataset

We checked if the dataset is balanced or not. For example, if we have 2000 positive reviews and 14000 negative reviews for training, our algorithm will not be able to learn the positive reviews properly. In our dataset, we had 20019 positive reviews and 19981 negative reviews. That means 50.1% positive reviews and 49.9% negative reviews. That means our dataset is already balanced.

B. Preprocessing Data

Data preprocessing is an important part of machine learning. If data is not processed correctly then it may misguide the process and show biased outcomes. In order to preprocess the data, we took the following steps:

- Null Value Checked
- Hyperlink Removed
- Line Break Removed
- Extra space Removed
- Removed Punctuation
- Removed Stopwords
- Lowered cases of all texts

V. Evaluation

We will implement the Decision tree, Random Forest, and Linear Regression ML algorithms to compare their performance. We will choose the algorithm that performs better than the others.

VI. Conclusion

We have only determined the algorithms that are to be implemented. Now we will need to see how they perform on our data set to reach our conclusion.