Exploring the Impact of Movie Reviews on Overall Success

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

***Abstract*—Movie reviews tell us whether a movie is enjoyed by the audience or not which may help to determine the box office collection 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 are essential to understand the impact of a movie on its audience. In commercial movies, the verdict of the movie is measured by its lifetime gross. It is essential for the movie industry to continue making movies. We may learn a lot about the overall insights by reading the movie reviews and analyzing the revenue alongside. Comments and opinions on a movie provide us with a more in-depth understanding of the movie's appeal to the audience which affects the movie revenue.

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.

Sentiment analysis is an excellent method for determining public opinion. In this method, we process the dataset to remove irrelevant information from the dataset. For instance, punctuation marks and stopwords do not tell us anything about the actual perspective of the audience and these can misguide the machine learning process. For these reasons we cleaned up the dataset and removed 'the', 'and', 'is', and punctuation marks. We additionally checked the nulled values, blank spaces, and new lines and removed them. Thus, the relevant data to the study is acquired for the next step phases which are:

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

II. Related Works

[9] 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 89.99% accuracy. Liang et al. (2015), and Hu et al. (2018) used sentiment and aspect-based sentiment analysis methods. These machine learning techniques enabled the opportunity to determine the sentiment of a text. It differentiates whether a text is positive or negative or neither. [10] [11]

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

Now, we will be discussing the whole process of our work starting from dataset organization to algorithm implementations. Here, we have applied necessary checkings to make the dataset ready to be trained with our machine learning models. The steps followed are shown in the "Fig: 1".

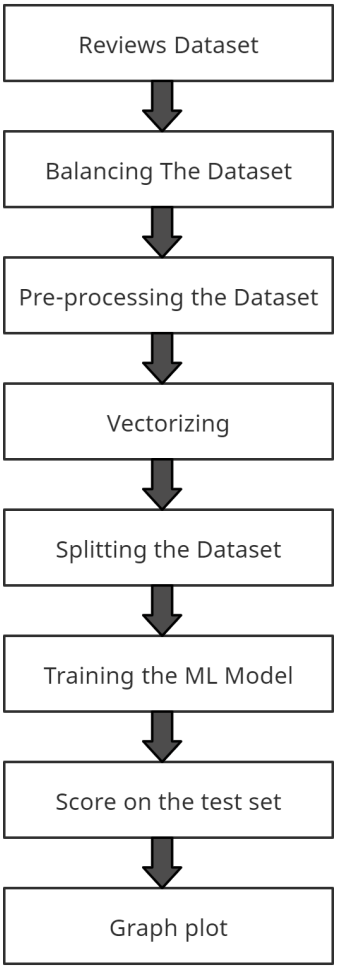

Fig. 1. Steps Followed

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

C. Training the ML Models

- LINEAR REGRESSION
  Linear Regression is considered to be one of the most important methods in statistical modeling. It allows us to quantify the correlation of a dependent variable and a group of independent or individual variables. It is possible that these linking variables affect two or more variables altogether. In this model, a linear correlation between variables is assumed. In linear regression, it is attempted to estimate the coefficients in the regression equation in order to lessen the gap between the predicted and observed values of the dependent variable. One method that may be used to get this is to make an estimate of the coefficient that will result in the variable that is dependent on the independent variable exhibiting the least degree of variation. The equation of the this regression has the dependent variable Y on left and explanatory variable X on right multiplied with the slope.Here, a is the intercept (value of y on x = 0). [17]

$$Y = a + bX \tag{1}$$

- RANDOM FOREST
  Random forest is a supervised machine learning algorithm used for classification and regression problems. Multiple models are combined into one in this algorithm. The following steps are used in the algorithm:
  - Choosing n number of random records from k number of records.
  - Building individual decision tree
  - Each decision tree will provide respective outputs
  - Final output will be based on majority voting or averaging.

  A large number of decision trees are generated in random forest in the training step. For classification, the majority of the trees are chosen. While for regression tasks, average prediction is chosen. [13]

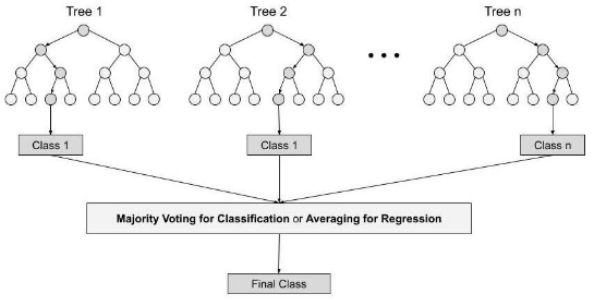

Fig. 2. Random Forest Algorithm [18]

Random forest generates a large number of decision trees while training. To classify, we choose the output of the majority of the trees. For regression tasks, we choose average prediction.

- DECISION TREE
  Another method to solve the classification problem is the decision tree. Even though the decision tree is not a regression but a popular supervised machine learning technique, it can solve the problems regarding regression. This model is a classifier with a tree structure. It is used for developing training models which can be used to derive data from the training data using simple decision rules to find out the class or value of the target variable. Here, a class label is predicted using a decision tree classifier. For that, at first, the tree's root is taken. Then it is compared. Through differentiation between the roots and the values that are in records, the next nodes are traversed depending on which branch node correlates to the respective value. [18].

V. Result

From the result of our experiment, it is seen that Linear Regression (LR) gives 10% accuracy, and Random Forest (RF) gives 65.93% accuracy. So from the results of our experiment, we can say that the Random Forest algorithm worked best in our research as it has given the highest accuracy. However, the accuracy level of the Decision Tree is not satisfactory. So we can conclude that Random Forest is the most cost-effective machine learning algorithm for predicting the box office revenue from reviews.

TABLE I
ACCURACY OF THE ML MODELS

Algorithm	Accuracy	f1-score
Linear Regression	0.10	0.10
Random Forest	0.6593	0.6593

VI. Conclusion

This article demonstrates our work on revenue analysis of IMDB Movie Reviews. Our studies' accuracy results reveal that Random Forest is the most effective machine learning algorithm for predicting the box office revenue of IMDB movies from reviews. We discovered which machine learning models are the most effective in predicting movie reviews as a result of this experiment. We can understand the impact of a movie on the audience and box office by analyzing the movie reviews.

References

[1] Kim, S., Park, N., Park, S. (2013). Exploring the Effects of Online Word of Mouth and Expert Reviews on Theatrical Movies' Box Office Success. Journal of Media Economics, 26(2), 98–114. https://doi.org/10.1080/08997764.2013.785551

[2] Thet, T. T., Na, J., Khoo, C. S. G. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. Journal of Information Science, 36(6), 823–848. https://doi.org/10.1177/0165551510388123

[3] Gadekallu, T. R., Soni, A., Sarkar, D., Kuruva, L. (2019). Application of Sentiment Analysis in Movie reviews. In Advances in business information systems and analytics book series (pp. 77–90). IGI Global. https://doi.org/10.4018/978-1-5225-4999-4.ch006

[4] Chirgaiya, S., Sukheja, D., Shrivastava, N., Rawat, R. (2021). Analysis of sentiment based movie reviews using machine learning techniques. Journal of Intelligent and Fuzzy Systems, 41(5), 5449–5456. https://doi.org/10.3233/jifs-189866

[5] Basuroy, S., Chatterjee, S., Ravid, S. A. (2003). How Critical are Critical Reviews? The Box Office Effects of Film Critics, Star Power, and Budgets. Journal of Marketing, 67(4), 103–117. https://doi.org/10.1509/jmkg.67.4.103.18692

[6] Gemser, G., Van Oostrum, M., Leenders, M. A. (2007). The impact of film reviews on the box office performance of art house versus mainstream motion pictures. Journal of Cultural Economics, 31(1), 43–63. https://doi.org/10.1007/s10824-006-9025-4

[7] Chintagunta, P. K., Gopinath, S., Venkataraman, S. (2010). The Effects of Online User Reviews on Movie Box Office Performance: Accounting for Sequential Rollout and Aggregation Across Local Markets. Marketing Science, 29(5), 944–957. https://doi.org/10.1287/mksc.1100.0572

[8] An improved sentiment analysis of online movie reviews based on clustering for box-office prediction. (2015, May 1). IEEE Conference Publication — IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/7148530

[9] (2018, July 30). Igi Global. Retrieved May 13, 2023, from https://www.igi-global.com/gateway/article/full-text-pdf/298652

[10] Hu, Y. H., Shiau, W. M., Shih, S. P., Chen, C. J. (2018). Considering online consumer reviews to predict movie box-office performance between the years 2009 and 2014 in the US. The Electronic Library, 36(6), 1010–1026. doi:10.1108/EL-02-2018-0040

[11] Liang, T. P., Li, X., Yang, C. T., Wang, M. (2015). What in consumer reviews affects the sales of mobile apps: A multifacet sentiment analysis approach. International Journal of Electronic Commerce, 20(2), 236–260. do i:10.1080/10864415.2016.1087823

[12] Basuroy, S., Chatterjee, S., Ravid, S. A. (2003). How Critical are Critical Reviews? The Box Office Effects of Film Critics, Star Power, and Budgets. Journal of Marketing, 67(4), 103–117. https://doi.org/10.1509/jmkg.67.4.103.18692

[13] The Science of Machine Learning. 2022. Random Forest — The Science of Machine Learning. [online] Available at: ¡https://www.ml-science.com/random-forest¿ [Accessed 17 August 2022].

[14] Medium. 2022. Twitter sentiment analysis using Logistic Regression. [online] Available at: ¡https://medium.com/nerd-for-tech/twitter-sentiment-analysis-using-logistic-regression-ff9944982c67¿ [Accessed 20 August 2022].

[15] Medium. 2022. Sentiment Analysis using SVM. [online] Available at: ¡https://medium.com/@vasista/sentiment-analysis-using-svm-338d418e3ff1¿ [Accessed 20 August 2022].

[16] Analytics Vidhya. 2022. Random Forest — Introduction to Random Forest Algorithm. [online] Available at: ¡https://www.analyticsvidhya.com/blog/2021/06/understanding-randomforest/

[17] About Linear Regression. (n.d.). IBM. Retrieved May 13, 2023, from https://www.ibm.com/topics/linear-regression

[18] Ali, J., Khan, R., Ahmad, N., Maqsood, I. (2012). Random Forests and Decision Trees. International Journal of Computer Science Issues, 9(5), 272-278.