PREDICTING BOX OFFICE SUCCESS OF MOVIES USING SOCIAL MEDIA

Midhun Josey
Undergraduate Student
Christ University (Deemed to be University),

Abstract

To know the fate of the movie before release is in the best interest of most of the film producers. Many numbers of factors affect the success of movies such as the cast, director, prequel or sequel, etc. In this system, we try to predict the success of a movie before the release by using data from social media. These social media sites mainly include Twitter, YouTube, and the IMDB database. This paper talks about how to use data mining, Support Vector Machine (SVM) and Neural Network Analysis to analyze the data and predict the success of a movie before its release. This system analyzes the sentiment from social media, the star value of the actor or actress and also the genre of the movie.

Keywords: Movie, Film, Success, Social media, SVM, NNA

I. Introduction

Lots of investment is being made into the film industry from various business sectors. The only motive they have is to get a huge amount of profit from the films. Thousands of movies get released across the world every year in different countries and in different languages. But due to several factors, some of the movies may not become successful some cannot make a profit. Even films with big cast or crew may not be a success at the box office. Thus, it is important for filmmakers to know the future of the movie before it gets released. With the help of data analytics, the success of the movie can be predicted efficiently. This is done with the help of data that we get from social media sites such as YouTube and Twitter and film databases such as the IMDB Database.

Movies are in fact a very large investment and a huge risk. Several factors come into play regarding the success of a film at the box office. The market value of the hero or heroine, their fanbase, the number of successful films made by the production company are all factors that come into play. Above all this the sentiments of the people matter. We get to know public opinion by looking into social media sites such as Twitter or YouTube. People express their sentiments in social media by posts or comments. Film studios upload their film's trailer on YouTube and people discuss it as comments or as posts in other social media platforms. In this way, social media plays a fundamental job in predicting the success of a film. Some studies have proven that Twitter data alone can be analyzed and used for prediction.

II. Literature Review

One paper [1] talks about predicting the revenue by the use of three layers. This system analyzes the value of movie stars, captures public opinions and analyze factors that can have an impact on the box office revenue. The three layers are the pre-processing task layer, the algorithm library layer and the visualization layer. The pre-processing task layer cleans and normalizes the data from various social media platforms and make it ready for analysis. The algorithm library layer analyzes the sentimental value, impact of cast and crew for a better prediction. The visualization layer provides a user-friendly system. It uses social Media Analysis System (SMAS) for analyzing social media posts and to give a better prediction.

Another system [2] focuses on twitter sentiments to predict the success of a movie. This system comprises of two modules which are the data collection module and the Predictive Engine. Data Collector retrieves information about movies from sources like IMDB, Wikipedia and social media including YouTube and Twitter. Twelve Features related to the movie are collected which is then categorized into conventional features and social media features. The prediction model calculates the rating and income by doing two separate sets of experiments. The ratings are predicted using Linear Regression Technique.

Another study [3] uses data from Twitter, YouTube, and the IMDB database. The factors include follower count on Twitter and Sentimental analysis of YouTube comments. This study follows three steps which are normalizing the data using the min-max method, K-Means clustering and Generating a predictive model. The movies were clustered into three classes Hit, Neutral and Flop. The predictive model uses Weka's J48 and Naïve Bayesian Algorithm to create two models validated by 3-fold cross-validation. It came to the conclusion that the popularity of a leading cast member is important for the success of a movie. The sentiment score and the number of views were found irrelevant in this scenario.

There is a system [4] that predicts movie success entirely on the sentimental analysis of tweets (opinion mining). This includes the following four steps: Data Collection, Data Pre-processing, Sentiment analysis, and Prediction. The tweets are collected using Twitter Application Programming Interface (APIs) and the following information is stored: Tweet Id, Username, Tweet text and Time of tweet. Data Pre-processing is done using distributed computing techniques and further cleans the data by regular expression matching. Sentiment analysis is done using the Lingpipe Sentiment analyzer and classifies the tweets as positive, negative, neutral and irrelevant. Prediction is based on statistics of the tweet's sentiments. We use the PT-NT ratio to predict the movie categories of the success and classify the movies as hit, flop or average. There is more than one factor that affects the movie box office but in this, they concentrate on the sentimental analysis of tweets.

Another paper [5] talks about the impact of movie trailers on the box office revenue. It concentrates on how people commit to a new film trailer and how people share it to other social media platforms and how it affects the box office performance. To collect the viewership data from YouTube they used Java-based Web-crawler and recovered the information by using YouTube's open APIs instead of HTML subpages. It uses 2SLS estimation with instrumental variables. It concludes that consumer engagement in a trailer is directly related to the activity of sharing the film trailer and directly affecting the box office performance.

A study [6] proposes box office predictions by analyzing online reviews. It proposes a decision support system for the movie investment sector using machine learning techniques. It takes data from the IMDB and by using Support Vector Machine (SVM), Neural Network and Natural Language Processing it predicts the success of a movie in the box office. This contains a variety of steps such as data acquisition, data cleaning, Feature Extraction, data integration, and transformation. The feature extraction uses about 15 features in this model. This will also take into consideration the release date, budget, number of screens in which it gets released, etc in determining the correct prediction. The entire phase consists of three sub-phases which are Sentiment Analysis, Support Vector Machine, and Neural Network Analysis.

Another paper [7] talks about data mining of online reviews and tweets for a better prediction of the movie box office. The public opinion of people can be understood correctly by analyzing online reviews. Mining of online reviews and tweets by using various Data Mining algorithms is the major part of this system. They have used the Sentiment Probabilistic Latent Semantic Analysis (S-PLSA) model for finding the sentimental analysis of reviews and tweets. For the sales prediction process, they have used Autoregressive Sentiment-Aware model (ARSA). Finally, P-N ratio is used to predict the movies as Hit, Average or Flop. It also says that we can use is Autoregressive sentiment and Quality Aware Model (ARSQA) for a better prediction and by also adding a quality factor.

III. Data Collection

This module is the most important as it recovers data about the films from various sources including film databases such as the IMDB and from Social media websites such as Twitter and YouTube. The data which we collect includes the number of followers on Twitter, the IMDB ratings, etc which are subjected to change. So, we must use the most recent data available. This is done with the help of YouTube and Twitter APIs other than relying on available movie datasets.

Features in Data

The data we collected from APIs are divided into different categories as follows:

Genre

Genres are framed by shows that change after some time as new genres are discovered and the use of old ones are ceased. Frequently, works fit into numerous sorts by method for acquiring and recombining these shows [8]. The IMDB Database contains a huge amount of information about any film at any particular point of time. The Films are categorized under genres such as action, adventures, animation, etc. The following table contains details about the rate weight relationship among genres.

Table 1. Rate Weight Relationship [3]

Genre	**Rate**
Action	0.52
Adventure	0.08
Animation	0.1
Comedy	0.02
Drama	0.06
Fiction	0.06
Romance	0.12
Thriller	

Follower Count of Actor or Actress

The number of followers on Twitter or Instagram (the greater value is taken) shows the popularity of the actor or actress in social media. We only consider the follower count of the top three members of the cast. If the actor or actress doesn't have a social media account it will put the movie to a disadvantage.

Number of views and comments

The popularity of a movie can be calculated by counting the number of views and comments on the official trailer or teaser of a movie has. People try to give their impressions about the trailer as comments. The more the number of views the more popular is the movie.

Number of likes and dislikes

Same as the number of views and comments, the number of likes and dislikes also matters on the official trailer. If the number of dislikes is more than a certain number in comparison with the number of views it means that the people don't want to watch the movie or they didn't like the trailer.

Successor

If the movie is a sequel or a prequel of a movie that was released already, then this movie has a higher priority in the prediction stage. The value 1 will be given it a successor else 0 is given. If the first part was a success then more people will wait for the movie than a movie without any part before it. It means more is the chance for the movie to be successful at the box office.

Sentimental value

The sentimental analysis must be done in two cases. First on all tweets related to the movies and second on all the comments on the official trailer of the movie on YouTube. The sentimental values are calculated with the help of Microsoft Power BI application which uses Text Analytic API [6]. This gives a sentimental value from 0 to 1 where value close to 1 means positive sentiment. The average sentimental value is calculated and multiplied with the number of tweets or comments. And this final value is considered as a feature.

IV. Prediction

The prediction process is done with the help of two parts. They are Support Vector Machine (SVM) and Neural Networks.

Support Vector Machine

We are implementing 10-fold cross-validation [6] in each of our experiments. In this, every component of the dataset is divided into ten groups. The first group will become the test data and the other nine will be training data. After the process is completed the second group will become the test data and others will be the training data. It will continue for all the 10 groups. SVM uses 4 kernels. They are Linear Kernel, Gaussian radial basis kernel, 3-degree polynomial kernel and Linear Support Vector Classifier Kernel (Linear SVC Kernel). Working with data more than two dimensions is a very difficult job in real life. Kernels are used in this type of situation when we are working with higher dimensional data. We have used several features as variables in this system and kernels can deal with limitless dimensional space. Scikit-Learn is used for the implementation of SVM [9].

Table 2. SVM Execution Comparison [6]

Kernel	**Exact**	**One Away**
Linear	56.16%	88.87%
RBF	55.36%	87.54%
Polynomial	52.58%	85.82%
Linear SVC	53.64%	85.43%

In Table 2, the exact and one away prediction accuracy for SVM can be seen. Out of all 4 kernels, all of them give a different result from one another. Among the four, RBF and Linear give a decent outcome. At the point where data is overlapping SVM is unfit to make hyperplanes properly. We know that 2D plotting is the best way to imagine and comprehend the vector areas and data relations [6].

Neural Networks

A Multi-Layer Perceptron neural system (MLP) [6] is utilized for the forecast. This MLP model is created utilizing Keras [10], an acclaimed python API for the neural system. Keras consecutive model is utilized to fabricate the model. Scikit-Learn [9] is additionally utilized for 10-overlap cross-approval. In the proposed model there are three shrouded layers, each has sixteen neurons. The info layer has fifteen hubs, and the last layer has five hubs for five outputs.

The three hidden layers of MLP gives a better outcome reliably by testing many numbers of hidden layers. Softmax and Relu functions are utilized in this model for final and hidden layers accordingly [6]. A typical issue seen in neural networks is overfitting. Dropout guidelines are embedded if overfitting happens. Thus, social media plays a vital role in predicting the success of a movie.

Table 3. Neural Network Execution Comparison [6]

Features	**Exact Match**	**One Away**
All Features	58.41%	89.27%

Neural networks and Support Vector Machine both give great outcomes, but neural networks create preferable expectation exactness over SVM. For careful expectation, neural system accomplished 58.41% exactness which is superior to past research works. On the off chance that we decrease the number of target classes, our expectation score will improve. The issue with SVM is, it experiences isolating information focuses effectively as a result of data isolation. All things considered, SVM is unfit to compute appropriate hyperplanes as it ends up mistook for frequent data isolation. In this specific circumstance, the neural system is performing better for arrangement and pattern recognition than other algorithms. Our MLP neural system will play much better on the off chance that we have more information in our grasp [6].

V. Conclusion

This study proves that features such as Genre, followers, view count, etc positively affect the success of a movie. It also proves that the sentimental analysis of tweets and comments are directly related to public opinion and plays a huge role in the success of a movie. We used algorithms such as Neural networks and Support Vector Machine for efficient prediction of the movie box office. Thus, the movie investors achieve a great deal with this kind of research and it helps them to efficiently invest in the movies that can be a success that can make a profit for their investment.

The following conclusions were made in this study:

- The sequel of a successful movie has a high chance of success than a normal movie in the same genre.
- The number of followers of the hero or heroine is important for the success of a movie
- The public opinion or sentiments of people on social media plays an important factor than all other factors combined.
- The study categorized the prediction in three categories: Hit, Neutral and Flop. It helps us to quickly analyze the fate of the movie and the investors can know how much return they will get by just looking into the prediction.

VI. Future Work

Further studies must be conducted using more features. The features such as budget, the number of screens the movie gets released, the value of the director, the success rate of the production company, etc can be included to get an improved result. This paper talks about two main social media sites: Twitter and YouTube. But if we include other platforms such as Facebook and Instagram the prediction accuracy may improve. The future work may also include predicting the exact collection the movie can make. This prediction can happen when we get to study the people's reactions for a few days after the movie gets released.

Bibliography

[1] Z. Liu, K. Chen, Y. Qu, S. Guo, C. Liu and C. Jia, "SMAS: An Investor-Oriented Social Media Analysis System for Movies," in *2018 IEEE International Conference on Big Data (Big Data)*, 2018.

[2] M. Ahmed, M. Jahangir, H. Afzal, A. Majeed and I. Sidiqqi, "Using Crowd-source based features from social media and Conventional features to predict the movies popularity," in *2015 IEEE International Conference on Smart City/SocialCom/SustainCom together with DataCom 2015 and SC2 2015*, 2015.

[3] K. Apala, M. Jose, S. Motnam, C. Chan, K. Liszka and F. Gregorio, "Prediction of Movies Box Office Performance Using Social Media," in *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2013.

[4] V. Jain, "Prediction of Movie Success using Sentiment Analysis of Tweets," *The International Journal of Soft Computing and Software Engineering [JSCSE]*, vol. 3, no. 3, pp. 308-313, 2013.

[5] S. Oh, J. Ahn and H. Baek, "Viewer Engagement in Movie Trailers and Box Office Revenue," in *2015 48th Hawaii International Conference on System Sciences*, Hawaii, 2015.

[6] N. Quader, M. Gani, D. Chaki and M. Ali, "A Machine Learning Approach to Predict Movie Box-Office Success," in *2017 20th International Conference of Computer and Information Technology (ICCIT)*, 2017.

[7] S. Magdum and J. Megha, "Mining Online Reviews and Tweets for Predicting Sales Performance and Success of Movies," in *International Conference on Intelligent Computing and Control Systems*, 2017.

[8] "Wikipedia," 24 April 2010. [Online]. Available: https://en.wikipedia.org/wiki/List_of_genres. [Accessed 23 May 2019].

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel and B. Thirion, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.

[10] Keras, "GitHub," [Online]. Available: https://github.com/keras-team/keras. [Accessed 23 May 2019].