*Journal of Information Systems & Information Technology (JISIT)* *Vol. 5 No. 2 2020*
*ISSN: 2478-0677* *72-80*

# Movie Success and Rating Prediction Using Data Mining Algorithms

Pirunthavi Sivakumar[1], Vithusia Puvaneswaren Rajeswaren[2], Kamalanathan Abishankar[3], E.M.U.W.J.B. Ekanayake[4], Yanusha Mehendran[5]

[1, 2, 3,4,5]Department of Computer Science and Informatics, Faculty of Applied Sciences, Uva Wellassa University of Sri Lanka

[1]pirunthavisiva08@gmail.com, [2]rvithusia@gmail.com, [3]findme994@gmail.com, [4]jayalath@uwu.ac.lk, [5]yanusha@uwu.ac.lk

**Abstract.** This project developed the models to predict the success and the ratings of a new movie before its release. Since the success of a movie is highly influenced by the actor, actress, director, music director and production company, those historical data were extracted from the Internet Movie Database (IMDb).The Box Office Mojo stores information about the cost of production of a movie and the total income of the movie. This information is helpful to determine whether the movie is successful or not in terms of revenue. A threshold was defined on revenue based on heuristics to categorize the movie into success or failure. Teasers' and trailers' comments were extracted from YouTube as those are very helpful to rate a movie. The keywords were extracted from the user reviews using a Natural Language Processing (NLP) technique and those reviews were categorized into positive or negative based on the sentimental analysis. A Random Forest Algorithm was trained using the features extracted from IMDb to predict the success of a movie. Further, the Naive Bayers model was trained using the user reviews extracted from YouTube to predict the rating of a movie. The models were tested on real datasets and the accuracy of those were evaluated respectively. Finally, two conclusions have been met that the rating of a new movie cannot be predicted in advance through the YouTube trailers' and teasers' comments and the success of a new movie can be predicted in advance by using the data or features collected from online. The performances of the models are decent enough compared to the existing models in the literature. The Success Prediction model can be used as an early assessment tool of movies since it has gained 70% overall accuracy and hence, useful for the people in the movie industry and the audience of the movies. YouTube allows to extract a limited number of user comments and hence, this factor could be negatively affected on the accuracy of the movie rating prediction. This abstract was presented at International Research Conference of Uva Wellassa University of Sri Lanka(IRCUWU2020).

**Keywords**: Data Mining, Natural Language Processing, Sentimental Analysis, Naïve Bayers, Random Forest



*Received: 30th Nov, 2020*
*Revised and Accepted: 29nd Dec, 2020*

# 1 Introduction

Watching movies have become one of the most popular entertainment factors in the 21st century. Many people want to invest in a movie. However, movies do not always see the face of success. Even though movie industry is profitable market, lots of producers and production companies face a massive loss. Success of movies is extremely complex matter because it's investment. Larger investment comes with larger risk.

The success and success rate of a movie is mainly dependent on people's perception. In today's digital world, generally people's opinion is in the form of reviews found online. The opening day and the first few weeks of a movie's release is very crucial, and hence production companies place a lot of importance on people's opinions and develop trailers and publicity strategies to get public opinion.

Our research topic is proposing a model for predicting the success and success rate of a movie. The excellent source of finding detailed information about almost every film ever made is through IMDb (Internet Movie Database) [10]. It contains a vast amount of data about general trends in films [10]. YouTube is also an excellent resource to get trailers' and teasers' user comments of films [4].

Data mining techniques allows us to predict the success of a future film before its release [11]. The main difficulty is to extract useful information from the IMDb and YouTube in the format of the source data. In this research, we have used customized dictionaries from the internet where different words that users commonly use in reviews will be grouped together and will be assigned a specific rate based on the admin's choice. By using Random Forest Algorithm we have classified the movie into hit or flop [11].

# 2 Methodology

We have collected data from IMDb (actor/actress, director, music director, Production Company) [11] and YouTube (teasers 'and trailers comments) [4] ,processed them (Data Pre-processing) ,transferred the processed data to train and test the model and predict the success and expected rating of the movie throughout our research time period.

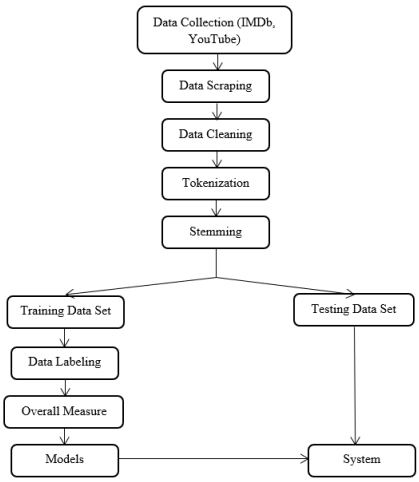


**Fig. 1.** Methodology Diagram

### 2.1 Data Collection

The initial dataset used is going to be collected from IMDb [11]. The following data types were obtained because they were containing the relevant information: Actor, Actress, Music Director, Director, Production Company, Release Date [11]. We collected all the important data manually for 200 movies which we have selected for our research project. We have selected movies which are correlated with some of the other movies for the purpose of having some connections/correlations of each other movies.

We used Python language to extract information about each video using the YouTube APIs. We collected 1000 comments per video because YouTube allows a maximum of 1000 comments per video to be accessed through the APIs. [7].

### 2.2 Data Pre-Processing

#### Data Scrapping

We used a tool "YouTube Scraper" for scraping comments from YouTube. This tool stores the extracted comments in CSV file (Comma Separated Value). But this tool is not that effective. This tool scraps all the comments from the video. As our PC does not have enough processing power and storage, it was very hard for further processing. Then we went with the YouTube API. We extracted user comments of films' trailers and teasers from YouTube using the YouTube API [7]. To scrape data from IMDb we searched for many online scraping tools which scrape data from IMDb. But we should pay for them to use those tools. So we got the data set from the Kaggle website, but there was a lot of data which we didn't need. So we manually collect some data which are having some correlations with each other to do the analysis process from that data set.

#### Data Cleaning

We cleaned the data in order to reduce the irrelevant attributes and select only the relevant attributes which would help in data analysis and the prediction process. We removed the punctuation and emojis from the user comments. After the Data Cleaning, we had robust dataset that avoids many of the most common pitfalls.

#### Data Tokenization

We split longer strings of text into smaller pieces, or tokens using Tokenization. Further processing was performed after a piece of text had been appropriately tokenized. We made sequence of strings into pieces like words, keywords, phrases other elements called tokens.

#### Stemming

By using data stemming we reduced different forms of a word and removed suffix, prefix in sentences.

### 2.3 Data Transfer

**Data set**: We got two types of data sets after data pre-processing.

- Training data set: This dataset is used to train the model; we have selected 200 movies to train the model.

- Test data set: The test data set is used to evaluate the performance of the model using some performance metric. It is important that no data from the training set are included in the test set. If the test set contains examples from the training set, it will be difficult to assess whether the algorithm has learned to generalize from the training set or has simply memorized it. We have selected 100 movies to test and validate the model.

### 2.4 Data Analysis and Prediction

**Data labeling:** We have labeled the words into two categories such as positive words (+), negative words (-).We got the positive, negative dataset from the internet. And according to those dataset we labeled the processed data into two categories (positive, negative words).So that we can calculate the percentage of positive features as well as the negative features.

**Overall measure:** We calculated the mean values by mathematical method. A user's comments may contain positive or negative features. So that we can calculate the percentage of positive features as well as the negative features. By this percentage, we can acquire a rating of a movie out of a hundred. By using this, we can gain the rating of a movie out of ten. Using Naive Bayes Algorithm we have calculated the mean values for the dataset collected from YouTube.

**Model:** Trained the model to predict the success of movies. We have selected 200 movies and using those dataset, we have trained the model. We tested the model to validate it. We have selected 100 movies and using those dataset, we have tested the model. Using Random Forest Algorithm, we have developed a model for the dataset collected to predict the movie success using from IMDb.

## 3 Implementation

### 3.1 Random Forest Algorithm

We used Random Forest Algorithm to develop a model for the dataset collected from IMDb [11]. Initially we started with the selection of random samples from the collected dataset. Next, this algorithm constructed a decision tree for every sample. Then it got the prediction result from every decision tree. Then the algorithm performed voting for every predicted result. Finally this algorithm selected the most voted prediction result as the final prediction result. Fig. 2 shows its work.
In Fig. 4, there are some fields which carry 0 (zero) as values. Zero means that field is an empty field. For example, if a movie does not have any main actor (hero) / if a movie is heroine oriented, we are assigning 0 to that field. We are assigning numerical values for all the features which are determining the success of a movie. These numerical values are stored in a CSV (Comma Separated Value) file. If we are not assigning any value and leave it as an empty field, the algorithm takes that space as an input value and trains the model as read in the CSV file.

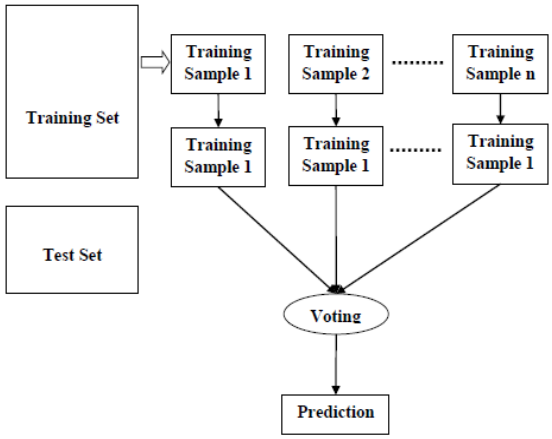


**Fig. 2.** Random Forest Algorithm Workflow

| Hero | Heroine | Director | Music Director | Production Company |
|---|---|---|---|---|
| Martin Lawrence - 1 | Margaret Avery - 2 | Malcolm D. Lee - 4 | David Newman- 3 | Universal Pictures - 5 |
| Kevin Hart - 6 | Tiffany Haddish - 7 | Brain Levant - 10 | Theodore Shapiro -26 | Gigantic Pictures - 8 |
| Jackie Chan - 12 | Regina Hall - 9 | David Frankel - 23 | Marco Beltrami - 30 | Amber Valletta - 11 |
| Robbie Amell - 14 | Amber Valletta - 11 | John Erick Dowdle - 28 | Christophe Beck - 31 | Cartoon Network Atlas Entertainme nt Telvan Productions Nine/8 Entertainme |
| Braeden Lemasters -19 | Kate Melton - 13 | Shawn Levy - 33 | Alan Silvestri - 38 | Hollywood Media Bridge Telvan Productions - 16 |
| Stacey Tavis -20 | Stacey Tavis - 18 | Anthony Russo - 35 | Marcelo Zarvos - 45 | Columbia Pictures Sony Pictures Releasing - 17 |
| Owen Wilson - 22 | Nia Long -21 | Stephen Chbosky - 48 | Rolfe Kent - 46 | Fox 2000 Pictures -25 |
| Justin Timberla ke - 51 | Jennifer Aniston - 24 | David Dobkin - 49 | Stephane Wrembel - 47 | Bold Films Brothers Dowdle - 29 |
| Brian Cox - 55 | Lake Bell - 27 | Woody Allen - 50 | Fernando Velázquez - 63 | Regency Enterprises - 34 |
| Javier Bardem - 57 | Rose Byrne - 32 | Andy Muschietti - 59 | Benjamin Wallfisch - 65 | 20th Century Fox - 39 |
| Nikolaj Coster-Waldau - 60 | Kate Hudson - 36 | David Fincher-69 | Alexandre Desplat-71 | Lionsgate Film - 40 |
| Alec Baldwin - 66 | Carla Gugino - 37 | Peter Jackson-72 | Howard Shore-74 | Tapestry Films - 41 |
| Brad Pitt-70 | Julia Roberts - 43 | Alejandro González Iñárritu-77 | The Dust Brothers-76 | gravier productions - 42 |
| Elijah Jordan Wood-73 | Rachel McAdams - 44 | Quentin Tarantino-79 | Gustavo Santaolalla-78 | Thema Production BBC films - 53 |
| Eddie Murphy-96 | Kate Winslet - 52 | Naomie Harris-85 | Michael Yezerski-83 | Mediapro Wild Bunch - 54 |
| Zac Efron-99 | Scarlett Johansson - 56 | Bruce Beresford-93 | Graeme Revell-86 | De Milo Productions - 62 |
| Reyan Renaolds-104 | Penélope Cruz - 58 | Joe Berlinger-97 | David Sardy-88 | Vertigo Entertainment - 64 |

**Fig. 3.** Data Set-1

| Hero | Heroine | Music Director | Director | Production | Output |
|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | h |
| 6 | 7 | 3 | 4 | 8 | h |
| 0 | 9 | 3 | 4 | 5 | h |
| 12 | 11 | 3 | 10 | 39 | h |
| 14 | 13 | 3 | 10 | 15 | h |
| 19 | 18 | 3 | 10 | 16 | f |
| 20 | 21 | 3 | 10 | 17 | h |
| 22 | 24 | 26 | 23 | 25 | h |
| 22 | 27 | 30 | 28 | 29 | h |
| 22 | 32 | 31 | 33 | 34 | h |
| 22 | 36 | 26 | 35 | 5 | h |
| 22 | 37 | 38 | 33 | 39 | h |
| 22 | 43 | 45 | 48 | 40 | h |
| 22 | 44 | 46 | 49 | 41 | h |
| 22 | 44 | 47 | 50 | 42 | h |
| 51 | 52 | 50 | 50 | 42 | f |
| 22 | 44 | 47 | 50 | 42 | h |
| 55 | 56 | 50 | 50 | 53 | h |
| 57 | 58 | 50 | 50 | 54 | h |
| 60 | 61 | 63 | 59 | 62 | h |

**Fig. 4.** Data Set-2

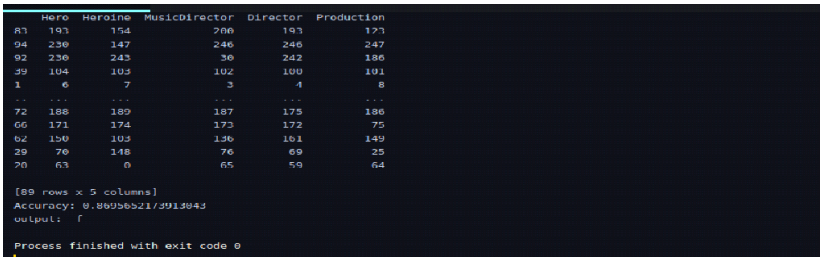


**Fig. 5.** Movie Success Prediction Sample Output

In Fig. 4, we put Hit as h and Flop as f. We got success/flop data from Box Office Mojo Website. Box Office Mojo is a website that tracks box office revenue in a systematic, algorithmic way. By using the revenue of a film, we labeled the film as Hit or Flop.

### 3.2 Naïve Bayer's Algorithm

Naive Bayer's uses an identical method to predict the probability of various classes supporting various attributes. This algorithm is usually utilized in text classification. We used this Algorithm to predict the expected rating of a movie by using YouTube Trailers' and teasers' comments. After pre-processing (Data Cleaning, Tokenization, and Stemming) and data-transfer the algorithm converted the dataset into a frequency table. Then it created a likelihood table by finding the possibilities. Fig. 6 shows its sample work. For instance the probability of positive words is 0.38 and probability of negative words is 0.61. Here we are giving it as a percentage out of 100.

| Words | positive/negative |
|---|---|
| Bad | negative |
| Fantastic | positive |
| worst | negative |
| Nice | positive |
| Super | positive |
| Exhausted | negative |
| Clumsy | negative |
| Harmful | negative |
| Hate | negative |
| Fantastic | positive |
| Admiring | positive |
| worst | negative |
| Harmful | negative |
| Beautiful | positive |
| Fantastic | positive |
| Bad | negative |
| Exhausted | negative |
| Nice | positive |
| Fantastic | positive |
| Super | positive |

| Freguency Table | | |
|---|---|---|
| Words | Positive | Negative |
| Bad | | 2 |
| Fantastic | 4 | |
| Worst | | 2 |
| Nice | 2 | |
| Super | 2 | |
| Exhausted | | 2 |
| Clumsy | | 1 |
| Harmful | | 2 |
| Hate | | 1 |
| Admiring | 1 | |
| Beautiful | 1 | |
| Grand total | 10 | 10 |

| Likelihood Table | | | |
|---|---|---|---|
| Words | Positive | Negative | |
| Bad | | 2 | 2/20=0.1 |
| Fantastic | 4 | | 4/20=0.2 |
| Worst | | 2 | 2/20=0.1 |
| Nice | 2 | | 2/20=0.1 |
| Super | 2 | | 2/20=0.1 |
| Exhausted | | 2 | 2/20=0.1 |
| Clumsy | | 1 | 1/20=0.05 |
| Harmful | | 2 | 2/20=0.1 |
| Hate | | 1 | 1/20=0.05 |
| Admiring | 1 | | 1/20=0.05 |
| Beautiful | 1 | | 1/20=0.05 |
| Grand total | 10 | 10 | |
| | 10/20=0.5 | 10/20=0.5 | |

**Fig. 6.** Naïve Bayer's Algorithm Example

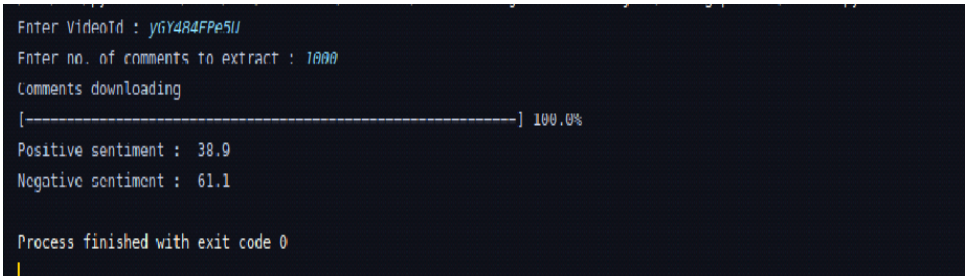


**Fig. 7.** YouTube Comment Analysis and Sample Output

Here we used YouTube API to extract the comments from YouTube. By giving the video IDs we extracted the comments of each and every film. Here we gave Video ID and number of comments which we wanted to extract as inputs and got positive and negative sentiment analysis percentage as output. YouTube API allows developers to access videos statistics and YouTube channels data via REST API call. YouTube allows a maximum of 1000 comments per video to be accessed through the APIs [7]. We did not get a small amount of comments for sentimental analysis, because the accuracy could be reduced when the number of comments is less. So we collected 1000 comments per video.

# 4 Testing and Evaluation

## 4.1 Rating Prediction

These days individuals will in general search for data and feelings to put together their own judgment with respect to. Verbal is viewed as a significant wellspring of data for the individuals [7]. These electronic verbs imply positive or negative proclamations made on YouTube comments. Regardless of whether it is a constructive or contrary comment, individuals will in general search for the comments to put together their own opinion with respect to. The majorities of the individuals give likes or dislikes dependent on their own suppositions. YouTube comments with more reactions will go to the highest point of the line. These days one negative remark can spread far and wide quickly [7]. Hence, that sort of comment will get a greater number of reactions than the other positive comments. So in the greater part of the recordings, negative comments are those which exist in the highest point of the line. At the point when we are scrapping trailers' or teasers' comments from YouTube, we can separate them from the top. We can't get every one of the comments, in light of the fact that YouTube permits a limit of 1000 comments for every video to be got through the APIs [7]. Comments with more reactions will be on the top and extracted through the APIs. Consequently most negative comments are extracted than the positive comments. Along these lines, more often than not the determined mean estimation of negative sentiment is greater than the determined mean estimation of positive sentiment.

**Table 1.** Rating Prediction Sample Outputs

| Movie Name | Expected Rating | Actual Rating |
|---|---|---|
| Avengers | 0.8 | 0.35 |
| Birth of the Dragon | 5.5 | 2.5 |

We created a dummy YouTube channel to check whether our developed model is performing accurately or not. Therefore we uploaded some videos to our YouTube channel and we asked some of our friends to put comments for the videos. Initially we asked them to put only negative comments, based on those comments our model calculated negative sentiment as 100% and positive sentiment as 0%. Fig. 8 illustrates the output.

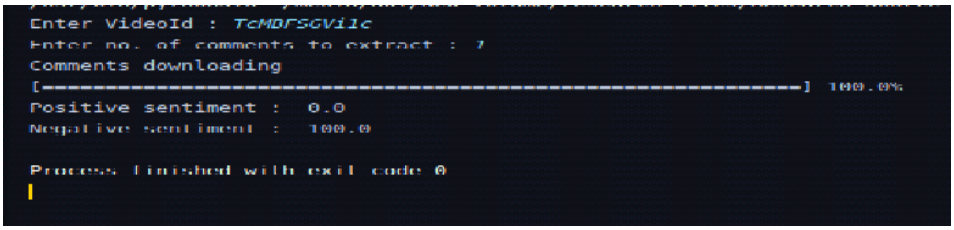


**Fig. 8.** YouTube Comment Analysis Sample Output_1

Then we asked half of them to put only negative comments and half of them to put only positive comments, based on those comments our model calculated negative sentiment as 50% and positive sentiment as 50%. Fig. 9 illustrates the output.

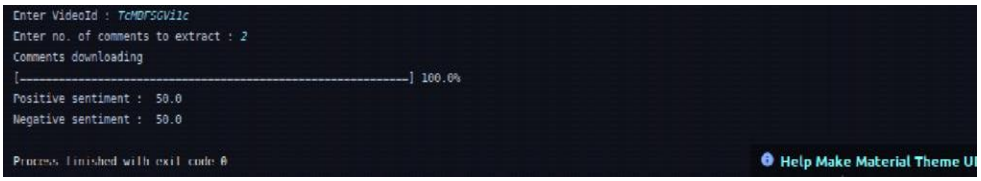


**Fig. 9.** YouTube Comment Analysis Sample Output_2

Then we asked them to put only positive comments, based on those comments our model calculated positive sentiment as 100% and negative sentiment as 0%. Fig. 10 illustrates the output.

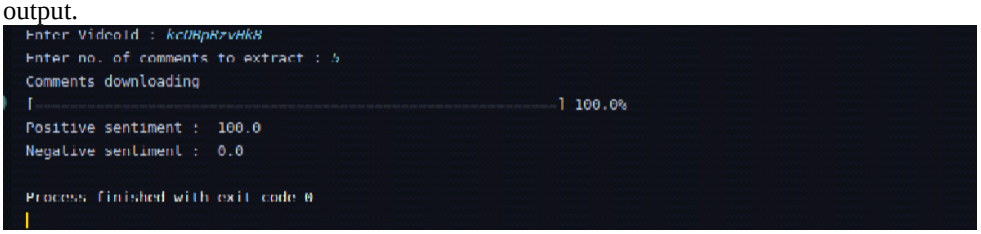


**Fig. 10.** YouTube Comment Analysis Sample Output_3

### 4.2 Success Prediction

We used 100 movies to test and validate the model. We used 200 movies to train the model. After training the model we used those test dataset to test the model. We put the data from the test dataset one by one to the trained model and check whether the model's output matches its real success.

We got success/flop details from Box Office Mojo Website. By using the revenue of a film, we labeled the film as Hit or Flop in the training data set.

**Table 2.** Success Prediction Table

| Movie Name | Expected Class | Actual Class | Results |
|---|---|---|---|
| The Yellow Birds | Flop | Flop | ✓ |
| Spoilers | Hit | Hit | ✓ |
| Fassbinder | Flop | Flop | ✓ |
| Wedding Crashers | Hit | Hit | ✓ |
| In the Heart of the Sea | Flop | Flop | ✓ |

## 5 Conclusions

We have finally come to a decision that we cannot predict the expected rating of a new movie in advance before the release through the YouTube comments of trailers and teasers. Because YouTube API extracts more negative comments than the positive comments as comments with more remarks will come to the top of the list. Mostly the output of the model gives more negative percentage than the positive percentage. Therefore, most of the time our model cannot give accurate results. Our developed model is working efficiently, but we cannot gain the expected rating of a new movie in advance before the release.

Next we have come to another decision that we can predict the success of a new movie in advance before the release by using the data or features collected from online. To reduce the complexity of dealing with the huge load of data we selected only up to 5 attributes, so as only the relevant data is used and can be understood by the user. We trained the model to predict the success of movies. When we were testing the model we got more than 70 % accuracy for all the testing data. So we came to the conclusion that our model is working properly and efficiently and predicts the success of a movie before its release.

## References


[1] Sivakumar, P., Rajeswaren, V.P., Abishankar, K., Ekanayake, E.M.U.W.J.B., Mehendran, Y. (2020). Proceedings of the International Research Conference of Uva Wellassa University. In: *International Research Conference - 2020*.

[2] Hu, X. and Aldous, D. (n.d.). Predicting Domestic Gross of Movies.

[3] Mestyán, M., Yasseri, T. and Kertész, J. (2013). Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data. PLoS ONE, 8(8), p.e71226

[4] Nanda, M., Pattnaik, C. and Lu, Q. (Steven) (2018). Innovation in social media strategy for movie success. Management Decision, 56(1), pp.233–251.

[5] cinemayward. (2015). Movie Rating System - cinemayward.

[6] Aggarwal, C.C., Chen, C. and Han, J. (2010). The Inverse Classification Problem. Journal of Computer Science and Technology, 25(3), pp.458–468.

[7] Krishna, A. (2014). Polarity trend analysis of public sentiment on YouTube.

[8] MH, S., S, W. and J, E. (2004). A data mining approach to analysis and prediction of movie ratings.

[9] Gothwal, K., Sankhe, D., Waghela, N., Sharma, M. and Yadav, R. (2018). Movie Success Prediction

[10] Pramod, S., Joshi, A. and A, G. (2017). Prediction of Movie Success for Real World Movie Data Sets.

[11] K, M., G, G., Agarwal, N. and Ghosh, I. (2018). A Data mining Technique for Analyzing and Predicting the success of Movie.