SocialMediaAnalytics - Movies

Vinay Rajagopalan , Sumanth Sripada , Nithesh Ramanna

2022-02-03

Project Objective

- Social Media is a very popular word and the role Social Media plays in Marketing domain is very crucial. In order to understand the above sentiment we decided to evaluate the role social media platforms like Twitter and YouTube play in predicting Rotten Tomatoes Audience score associated with a movie and we also tried analyzing the success of a movie in terms of opening weekend revenue. This project required us to do a deep dive into multiple factors associated with a movie in order to do this. There were three parts to this project.
- In the first part we collected data for all movies released in 2021. We used IMDB, Wikipedia and Rotten Tomatoes as source of information for this purpose.We made use of web scrapping primarily to obtain the cast, meta score, opening weekend revenue, production house and other attributes associated with a movie.
- The second part was analyzing twitter sentiment associated with a movie prior to movie being released. In our case we have used Tweets from 30 days prior to the movie release date. We have used these tweets to predict what might be the Rotten Tomatoes audience score for movie after release.
- In the third part we wanted to evaluate what plays a key role in movies opening weekend revenue. We gathered all the statistics associated with movie cast like the average revenue they generated, average total movie run time for the cast, average meta scores their movies received and some more features.We took these statistics for the last 5 movies for a every cast associated with a movie. In addition various key parameters on Social Media platforms like twitter and youtube to understand user engagement and following a movie twitter screen name had in the last 30 days prior to release. It also involved gathering similar kind of statistics from youtube trailers and teaser videos.
- In conclusion we were able to predict the Rotten Tomatoes Audience Score for a movie in the scale of 1-5 based on the 30 days Tweets for the movie prior to the release date. We were also able to see that engagement scores on Social Media platform could be a major factor in helping us predict how a movie would perform in terms of opening weekend revenue.

Fetching the Data

- All the required Movies list and their respective statistics were fetched from IMDB using the rapid api and web scraping. We have individual scripts that scrape for cast, cast performances, movies and Rotten Tomatoes reviews
- We gathered Historic Tweets from twitter for each movie 30 days prior to the movie release date.
- Youtube statistics for user engagement for release and trailer videos were gathered using google APIs.

Data Collection:

The data was gathered from multiple sources like IMDB, Wikipedia, Rotten Tomatoes, Twitter and Youtube.

- We created a Movie database of all movies released in 2021. There were a total of 357 movie.We gathered the IMDB ids for these movies using rapid API which would be the primary key for further tables that were created.
- We created table with movie attributes like total run time, opening weekend revenue, release date, meta score, meta critics count. We scarped IMDB to obtain these values for all movies in our movie table.
- We then scarped IMDB to obtain all the cast associated with a movie. We also scraped IMDB for obtaining the last 5 movies prior to the movie in context that these actors acted in and obtained all the movie attributes fo those movies.
- We used twitter API to search for movie screen name and collected all the historic tweets associated with the screenname 30 days prior to the movie release date.We also obtained key customer engagement parameters like reply, retweet count,quote count and like count along with these tweets for this period.
- In addition we also collected YouTube metrics for movie release or trailer videos that were released prior to movie release date.
- Lastly we collected 10 user reviews for Movies released from 2011 to 2017 along with the ratings provided by the user from rottentomatoes. We collected around 4200 ratings for this purpose.

Preprocessing

Text:

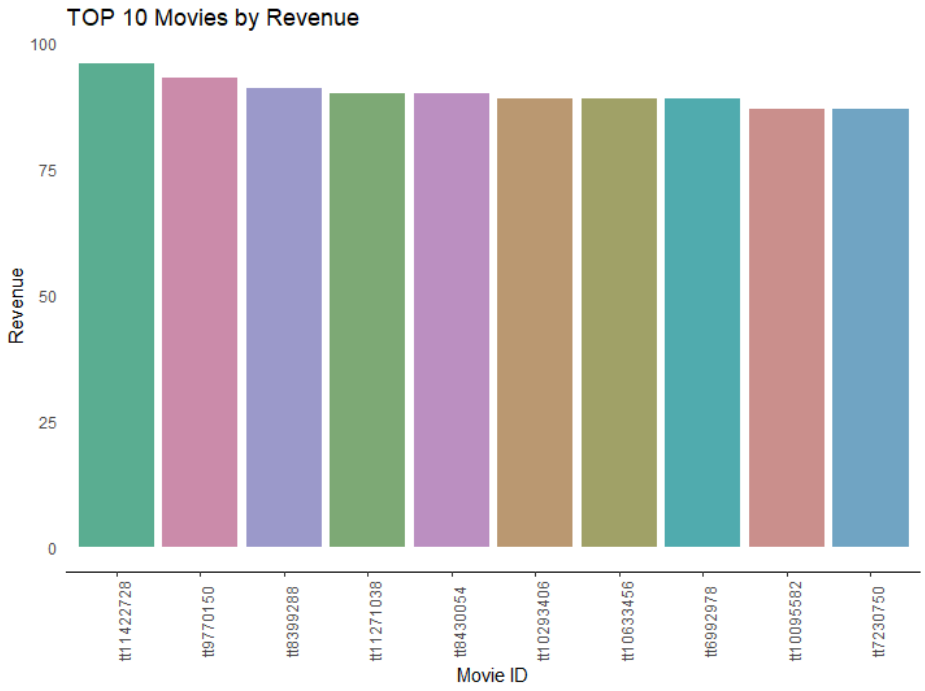
- Removing Punctuation and Special Characters
- Removed Special Characters from the text
- Tokenization to extract the words from the text
- Removed the Stopwords
- Lemmatization to extract the lemma of the words.

Numeric Columns:

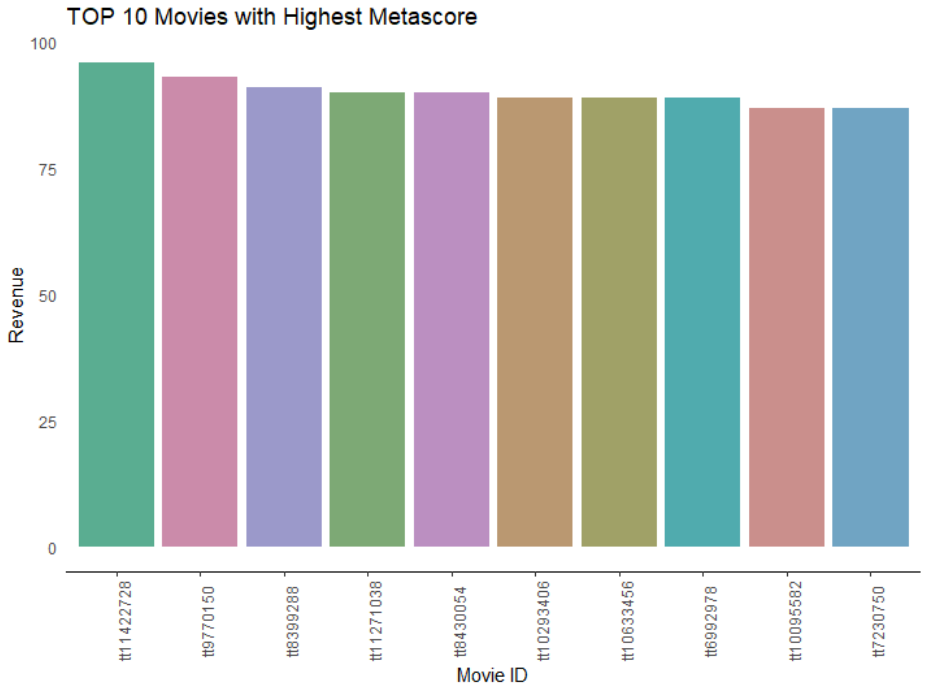
- All NAS were filled with 0
- The data is scaled

INSIGHTS:

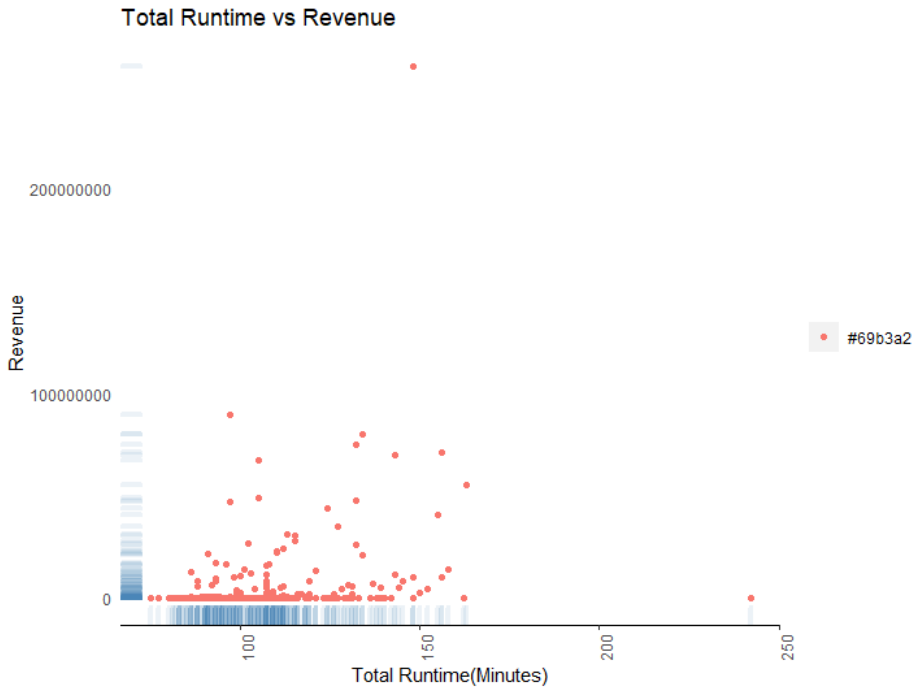
Top 10 Movies by Revenue :



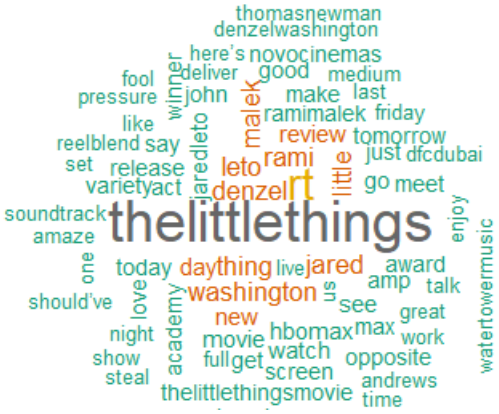
Top 10 Movies by Metascore:



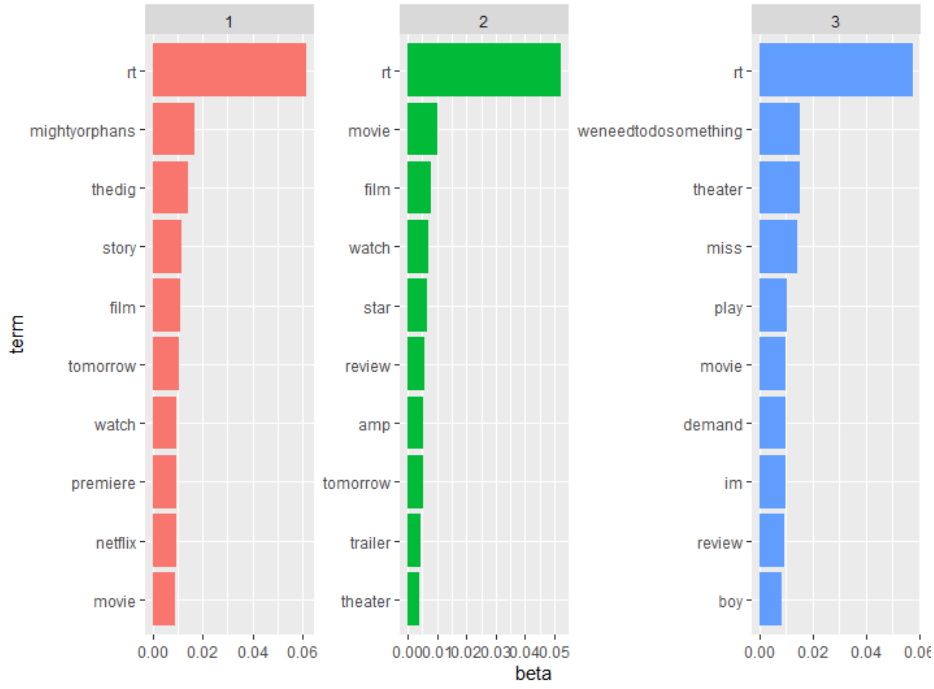
IMDB Movie runtime vs Revenue:



Word cloud for one Movie:



Top terms extracted by topic Modelling:



MODEL Train Test Data

Model 1 : Predict Rotten Tomatoes Audience Score in scale 1:5 based on historic tweets from movie screen name

We performed all the Text preprocessing steps for historic twitter tweets and rotten tomatoes reviews we collected. In addition to the text we also collected sentiment score for reviews and tweets based on Bing, AFinn, Loughran and NRC dictionary. We also did a parts of speech evaluation for each review and tweets. These columns were used be used as additional feature along with the text from tweets and reviews. We used Rotten Tomatoes reviews text along with the sentiment and POS in order to train and test our model. Once the model was trained we used our model to rate the tweets we had collected on a scale on 1:5 and lastly aggregated the results to obtain the audience score for the movie associated.

Model 2: Predict Opening Weekend Revenue Generated by the movie based on statistics of Cast and User Engagement of the movie.

In order to train our model we used the statistics collected from various sources associated with cast of the movie like meta score, opening weekend revenue, meta score, meta critics count for all the cast and took an average per movie. We also provided used various metrics of user engagement like view count, like count, comment count, retweet count, reply count, qoute count and favourite count. All these labels acted as the feature for the model. Opening weekend revenue was the independent variable. The data was then split into 60 40 split to train and test the model.

Model Evaluation

Model 1:

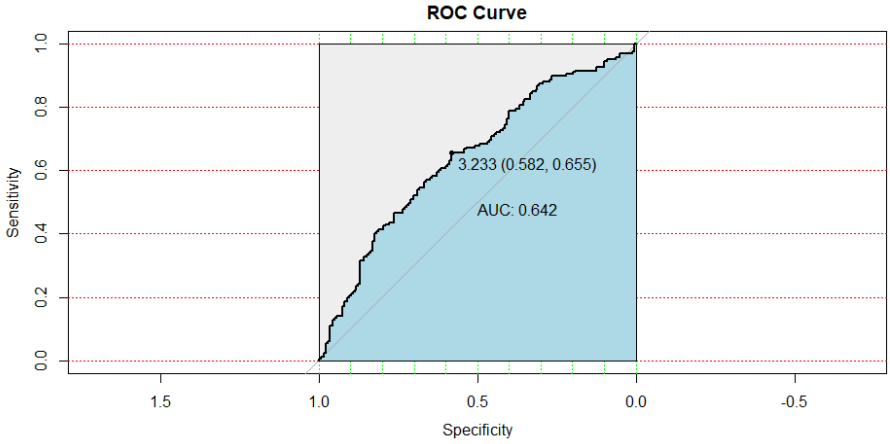
Residuals: Min 1Q Median 3Q Max -3.8091 -0.7861 0.1778 0.9461 5.0429

Coefficients: (1 not defined because of singularities) Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.414556 0.029746 114.789 < 2e-16 ***ADJ -0.019355 0.010533 -1.838 0.066225 .***
ADP 0.044469 0.053055 0.838 0.401997
ADV -0.004371 0.012517 -0.349 0.726953
AUX -0.001612 0.048105 -0.034 0.973275
CCONJ -0.109638 0.115862 -0.946 0.344076
DET -0.066403 0.073991 -0.897 0.369548
INTJ -0.036958 0.020008 -1.847 0.064818 .
NOUN 0.007407 0.004342 1.706 0.088136 .
NUM 0.006308 0.057898 0.109 0.913243
PART 0.006408 0.059270 0.108 0.913907
PRON -0.061259 0.037474 -1.635 0.102203
PROPN -0.187627 0.054238 -3.459 0.000549 SCONJ 0.001930 0.114139 0.017 0.986513
SYM -0.064740 0.123906 -0.522 0.601365
VERB -0.014559 0.009868 -1.475 0.140180
negative.x -0.071344 0.017815 -4.005 6.35e-05 ***positive.x 0.078514 0.018312 4.287 1.86e-05*** sentiment_Bing NA NA NA NA
Sentiment_affin 0.035320 0.006099 5.791 7.65e-09 ***constraining 0.085653 0.111939 0.765 0.444222***
litigious 0.229709 0.080810 2.843 0.004503 *negative.y 0.010872 0.024279 0.448 0.654329*
positive.y 0.028711 0.027359 1.049 0.294056
superfluous 0.084067 0.225905 0.372 0.709816
uncertainty -0.012642 0.040510 -0.312 0.755001
anger 0.083985 0.029483 2.849 0.004419 ***anticipation -0.033272 0.018555 -1.793 0.073052 .***
disgust -0.265395 0.030009 -8.844 < 2e-16 fear 0.067520 0.022613 2.986 0.002849 ** joy 0.063354 0.023764 2.666 0.007715 ** negative 0.002388 0.022471 0.106 0.915387
positive -0.004233 0.015742 -0.269 0.788030
sadness 0.051774 0.024589 2.106 0.035317 *
surprise 0.048191 0.022166 2.174 0.029769 *
trust -0.041033 0.017940 -2.287 0.022247 *
V1 -12.374936 5.947702 -2.081 0.037547 *
V2 9.018240 3.365201 2.680 0.007403 ** V3 2.854665 2.044003 1.397 0.162628
V4 2.257162 1.688442 1.337 0.181372
V5 0.780262 1.691585 0.461 0.644642
V6 -8.565854 1.933365 -4.431 9.71e-06 ***V7 -0.862926 1.759385 -0.490 0.623834***
V8 0.044520 1.786201 0.025 0.980117
V9 -5.396904 1.528272 -3.531 0.000419 V10 1.616184 1.588527 1.017 0.309034
V11 0.210760 1.530892 0.138 0.890509
V12 3.389227 1.560653 2.172 0.029953 *
V13 -2.318960 1.500041 -1.546 0.122219
V14 -2.209275 1.383606 -1.597 0.110419
V15 -3.027178 1.469192 -2.060 0.039437 *
V16 -5.420973 1.421766 -3.813 0.000140 ***V17 2.486140 1.454595 1.709 0.087517 .***
V18 -4.634152 1.445167 -3.207 0.001356 *V19 -5.214543 1.376451 -3.788 0.000154* ** V20 3.463930 1.358374 2.550 0.010816 *
— Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.194 on 3236 degrees of freedom Multiple R-squared: 0.2061, Adjusted R-squared: 0.1928 F-statistic: 15.55 on 54 and 3236 DF, p-value: < 2.2e-16

- Based on the F-statistic score we can see that we clearly see a relationship between sentiment of the tweets prior to movie release and Rotten Tomatoes Audience score for a movie. The model performance could be improved by providing more training data.

The AUC and confusion matrix for the same is as follows:




```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4   5
         1   1   3  58 111   4
         2   1   4  24 127   8
         3   0   1  15 232  39
         4   0   0  11 294  88
         5   0   1  11 269  84

Overall Statistics

               Accuracy : 0.2872
```

Model 2:

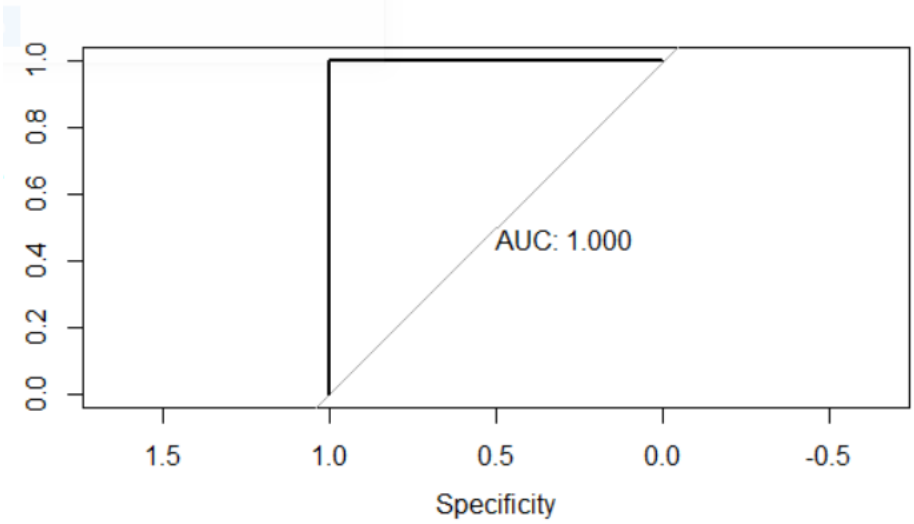
Here in model 2 we considered the combination of below features based on the significance of these features. The features are: cast_last5_movies_avg_meta_score, cast_last5_movies_avg_meta_critics_count, cast_last5_movies_avg_revenue, cast_last5_movies_avg_total_runtime, agg_reply_count,total_runtime_in_minute, agg_retweet_count, agg_favorite_count,agg_quote_count

Residuals: Min 1Q Median 3Q Max -0.3205 0.0000 0.0000 0.0000 0.2636

Residual standard error: 0.2726 on 8 degrees of freedom Multiple R-squared: 0.9853, Adjusted R-squared: 0.8238 F-statistic: 6.101 on 88 and 8 DF, p-value: 0.005014

From the F-statistic score we can see the relationship between the casts' statistics on previous movies and their interactions in social media with movie opening weekend revenue.

The AUC for the same is as follows:



Model Interpretation

Model 1:

We were able to see that there is a relation between sentiment of tweets in last the 30 days prior to the release date to the Rotten Tomatoes Audience Score of the movie.

Model 2:

Cast Previous Performance and User Engagement both play a crucial role in evaluating opening weekend revenue for a movie. When we trained the model with interactions between these features we obtained the best results thus proving our hypothesis that user engagement on social media platforms helps in driving the opening weekend revenue for a movie.

Shiny Application

- To understand the insights better, a shiny application was built and has individual tabs for IMDB, Twitter and YouTube for quick summary statistics.
- Wordcloud for each movies was included in a separate tab

Challenges Faced

- There were many constraints we faced during data collection. Twitter restricts the number of historic tweets we could retrieve thus making it difficult for us to collect more data. For our use case we could only collect 100 tweets per screen name in the 30 days period prior to a release date.
- The second challenge we faced was with Training Data for the first model as we were only able to get 10 reviews per movie with valid ratings from Rotten Tomatoes our training data for the first model only included 4000 reviews and we feel increasing the test data size could be beneficial in improving the accuracy of our model.
- For our second model out of the 357 movies we tried collecting data for 194 movies had the required independent variables available for training and testing our model. We feel if we could collect data for more movies from previous year we will be able to obtain better results.

Future Work

Based on the challenges we faced and constraints in terms of API limits we feel we feel we would be able to improve on our models performance if we could collect more data and add features that involved actors and production houses twitter and YouTube channels. Due to the time constraint we were unable to deep dive further into these aspects and would be a great place to start again. For the first model we feel that evaluating the sentiment of comments on release and trailer videos on YouTube could also help improving our model.Model 2 Currently has an AUC of 1 and we were not able to validate whether the AUC is accurate or not. We plan to look into this as a future work.