Analysis of Key Factors Influencing TV Shows' Success

Aakarsha LNU
alnu2@binghamton.edu
SUNY Binghamton
Binghamton, New York, USA

Riya Yeshwant Thakur
rthakur1@binghamton.edu
SUNY Binghamton
Binghamton, New York, USA

Aayushi Ahlawat
aahlawa1@binghamton.edu
SUNY Binghamton
Binghamton, New York, USA

Ishan Bagchi
ibagchi1@binghamton.edu
SUNY Binghamton
Binghamton, New York, USA

Tarun Tiwari
ttiwari1@binghamton.edu
SUNY Binghamton
Binghamton, New York, USA

ABSTRACT

These days, television and movies are the two main forms of entertainment. Everybody enjoys watching and reading fascinating stories. When we watch television, we observe that many people discuss their experiences. On social media, anyone can publicly voice their opinions on any subject. First off, the big social networking sites like Twitter, Instagram, TMDB, and Reddit create a ton of information about user behaviour's, news, rising trends, etc. There is a huge amount of user-generated content available, and this alters the environment significantly. The genre, cast, number of seasons, number of episodes, production companies, networks, episode run time all play a role in determining whether a television program is successful or not. Thus, by examining these elements for a range of programs, we aim to predict which TV show has a higher chance of becoming successful and displaying the live results on the dashboard.

KEYWORDS

Tv shows, Genre, Cast, Budget, Reviews, Region, Twitter, Reddit, TMDB

1 INTRODUCTION

We all watch a lot of television. Who does not enjoy having some spare time on the weekend? With the increasing popularity of OTT platforms and the media business, individuals have more options for different TV series and movies, which has increased competition in the media sector. People in the entertainment industry are always seeking for new strategies to make their series or movie a success. Thus, data analysis is becoming increasingly important in managerial decision-making in media enterprises throughout the world. The dependence of subscription video on demand providers, such as Netflix, on data analytics to support choices regarding new content investment is widely documented.

Technology advancements and the industrialization of television have an impact on how TV programs are produced, which is why it is moving in a variety of directions. The expanding entertainment and commercialization of TV series, as well as the marketization and industrialization of the TV drama industry chain, have all contributed significantly to the rise in TV series ratings. There are many researches going in this field which is also proposed in **[1]** An Analysis of the Factors Affecting the Ratings of TV Series and which further states that politics, economy, and technology are dynamically influencing the popularity of tv shows. These three forces inhibit and influence one another, resulting in a complex social backdrop that supports the expansion of TV series and the rise in TV series ratings. Also, more examples of such researches can be seen in **[2]** Predicting the performance of TV series through textual and network analysis: The case of Big Bang Theory where most of the script-related factors were considered which includes dialogue-related (e.g., length, language complexity, sentiment), while viewer reviews (i.e., the quantity of reviews as a popularity indicator and viewer ratings as a measure of appreciation) are used to gauge performance.

For this project, we have created an interactive dashboard that will be updated in real-time, showing results for the three research questions/areas based on our datasets and the type of data we have gathered. Our research objectives shall revolve around the factors influencing the popularity of TV shows based on data collected from TMDB, Twitter and Reddit.

2 WHAT IS TMDB?

TMDB **[3]** is a website where users may discover millions of movies and TV episodes. It gives us the option to look for the most well-liked films or TV shows that are now being seen in cinemas, on TV, or that are available for rental. Users of the portal can also publish reviews of movies and TV series and rate them. The website offers suggestions depending on the movies the user has looked up or seen. The TMDB API, which is accessible on the TMDB API Developers page, will be used. The purpose of this project is to evaluate the TV show's performance based on important elements like budget, cast, etc.

Twitter is an open social network that people use to converse with each other in short messages, known as tweets. People publish their comments regarding actors, production companies, and TV shows, which is useful for our study of the most popular and trending shows and their dependence on various factors.

Reddit **[4]** is a social news aggregation, content review, and debate platform. The website's material, which includes links, text entries, photographs, and videos, is contributed by registered users and is then rated by other users. Posts are arranged into user-made boards called "communities" or "subreddits" according to their subjects. When there are enough upvotes, posts that have received the most upvotes will eventually appear on the first page of the website.

3 DATA SET

TMDB, Twitter, and Reddit are the datasets that have been considered for collecting, analysis and visualization of the data for gathering 4728 TV shows and over 1000 distinct TV series changes every day, including changes to their ratings,show information, number of episodes etc.

3.1 Collection Methodology

We were successful in gathering pertinent data from the TMDB, Twitter, and Reddit platforms as part of the prior subproject. We had gathered information from all the aforementioned platforms and filtered it as needed to create a viable data frame for analysis which is described below.

TMDB:
TV Shows were collected in descending order with respect to "first air date" using GET discover API and were stored in tmdb collection. We obtained a list of newly released TV shows using the tmdb database we previously constructed, and for each one, we extract the list of reviews produced and add the TV ID linked with it to our "reviews" collection.

Twitter:
The stream of tweets was compiled using Python programming. The Twitter Streaming API, **[5]** resource URL: https://api.twitter.com/2/tweets/sample/stream, was used to collect real-time tweets. The URL had several additional factors that were useful for the study that follows for our project.

Reddit:
Python script was built which was used to generate the subreddit collection, with the help of temporary OAuth token from Reddit. Then, in order to read the data and stream new postings, an API was established. The most recent post was used to read or streamed using a variety of sub-reddits as parameters, producing real-time data.

4 METHODOLOGY FOR DATA ANALYSIS

4.1 TMDB

- For TMDB, firstly a connection is established with MongoDB to retrieve and start with the pre-processing of the available data.
- In the formed dataframe, the dependent and independent variables were as follows:
  Independent Variable: id,name,first_air_date,genre,cast_details,episode_run_time,number_of_episodes,number_of_seasons,overview, production_companies,status,type
  Dependent Variable: popularity
- After formation of dataframe, pre-processing is done.
- After the pre-processing, model training and testing was done for the data frame which contains 58 columns.
- Data frame was divided into training and testing sets and supervised learning modelling was used in which a training set is used to instruct models to produce the desired results.
- After which graphs are plotted to predict the popularity based on other independent factors.

4.2 Twitter

- For twitter dataset, tweets related to TV shows were fetched from the data stored in MongoDB followed by storing it in pickle file.
- Descriptive Analysis was performed on the retrieved data to identify various patterns followed in the dataset and a time series graph was plotted between the tweet count and TV Shows tweet frequency count for a particular duration.
- Pre-processing of tweets was done using a pre-processor library in python.
- Popular TV Shows were fetched using internal API of TMDB i.e., https://api.themoviedb.org/3/.
- Further, various analaysis such as finding top 100 TV shows to determine their hype on twitter was done.
- And lastly the result was visualized graphically.

4.3 Reddit

- For each of the subreddit data stored Top 3,4 phrases were retrieved followed by formation of data frame.
- Further for descriptive analysis, number of comments for each of the subreddits was retrieved and also, analysis for the number of comments for the top 5 subreddit.
- Lastly, data analysis was done for the politics subreddit and a time series was plotted to for determining the comments on a particular date.

5 METHODOLOGY FOR CREATING DASHBOARD

For creating live data dashboard, there are several tools, library packages, web applications and software's available. From our data analysis project, we have answered and visualized the results of all our research questions that were:

1. How are elements like the genre, cast, number of seasons, number of episodes, production firms, networks, and length of the episodes affecting the success of TV shows?

- From the TMDB data analysis it could be figured out that top 5 features that most affect the success of TV Shows are : returning series status, plot status, ended features status, canceled feature status and in production status with impact of greater than 2.0.And the least important feature is genre erotic with an impact of approximately 0.01.

2. To what extent are popular TV series on TMDB hyped on Twitter, and if this hype is positive or negative?

- On the basis of tweet frequency top 20 TV Shows were retrieved to calculate the polarity of each of the show. After calculating the polarity score hype was marked as positive, negative or neutral hype. On the basis of analysis made positive and negative hypes were in the range of 0-2000 with highest positive hype at 6000 and negative hype at 2000 respectively.

3. What phrases/terms from major TV series are becoming popular and most used or admired by the public on their respective subreddits?

- From the subreddit analysis 5 major phrases which are becoming popular are: BigBrother, thewalkingdead, HouseOfCards, brooklynninenine, television.

5.1 LIVE DASHBOARD

- For this project, we have created an interactive dashboard that is getting updated in real-time, showing results for the above-mentioned research questions.
- The dashboard is implemented using the Flask framework, and all the plots and graphs are built using various Python libraries.

5.2 DASH LIBRARY

- The original low-code framework, Dash, allows users to create data apps quickly in Python, R, Julia, and F (experimental).
- Dash is the best tool for creating and delivering data apps with unique user interfaces because it was built on top of Plotly.js and React.js. It is especially appropriate for anyone who handles data.
- Dash abstracts away all the technologies and protocols necessary to create a full-stack web app with interactive data visualization using a few straightforward principles.
- Dash applications are displayed on a web browser. Your apps can be installed on virtual machines (VMs) or Kubernetes clusters, and then shared using URLs. Dash apps are inherently cross-platform and mobile-ready because they are viewed in a web browser.
- Dash is an open source library that was made available under the MIT license. Dash is created by Plotly, which also provides a platform for creating and deploying Dash apps in an enterprise setting.

5.3 Implementation

- The Dash HTML Components module (dash.html) has a component for every HTML tag as well as keyword arguments for all of the HTML arguments.
- TMDB Dataframe is converted into a csv file named predictions.csv using which a layout have been using DASH library to display the graph for Popularity Predictions for the Un-Released TV Shows.
- For reflecting the twitter data, dataframe made for twitter is converted into showspolarity.csv, showing the analysis of Hype of the most popular TMDB TV shows on Twitter.
- Similarly for reddit, after converting dataframe into csv files named class.csv and datecom.csv, DASH library is used to show the graph for Number of Posts by each subreddit and a time series graph.
- Below are the results of our implementation:
- **Figure 1:** shows a graph plot between no of posts by subreddit (y-axis) vs subreddit (x-axis) i.e Number of posts by each sub-reddit is displayed.
- **Figure 2:** shows a graph plot between various values (y-axis) and show name (x-axis) i.e it is analyzing hype for the most popular TMDB TV Show on Twitter.

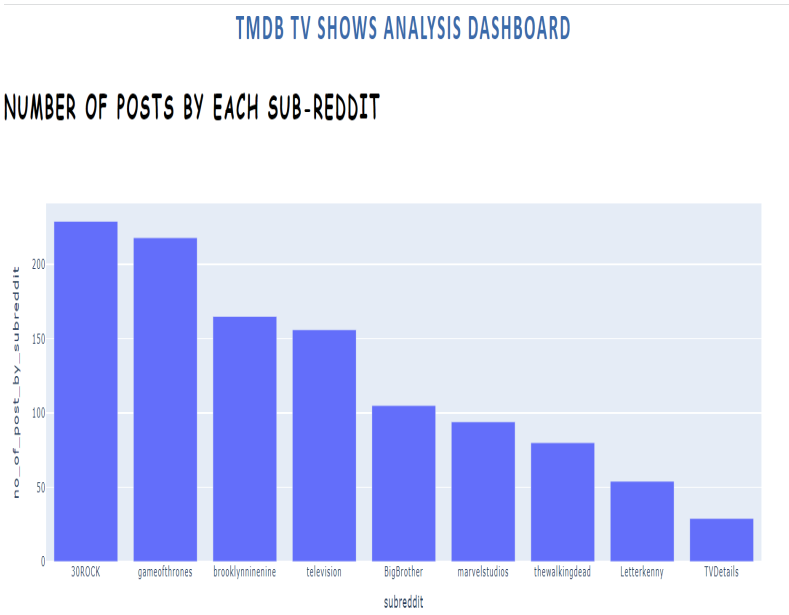

Figure 1: Number of Posts by each sub-reddit

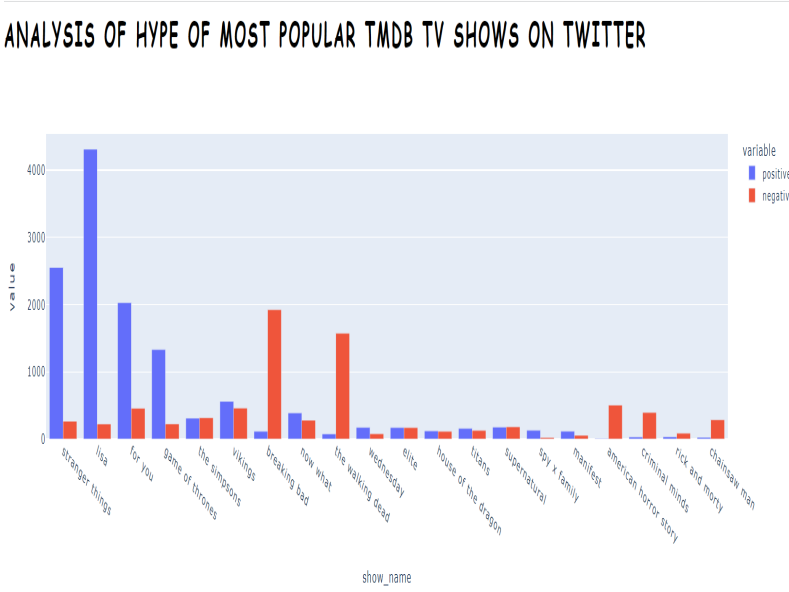

Figure 2: Analysis of Hype of Most popular TMDB TV show on Twitter

- **Figure 3:** shows a table for the popularity prediction of future TV Shows and displaying attributes are ID, Name and Predicted Popularity.
- **Figure 4:** plots a time series graph for the politics subreddit between no of comments (y-axis) and a specfic time durations (x-axis).

POPULARITY PREDICTIONS OF FUTURE TV SHOWS

ID	NAME	PREDICTED_POPULARITY
0	Western	1
1	The Buccaneers	0.180592021
2	Still Up	0.291712393
3	The Gallows Pole	0.644812522
4	Mary & George	0.051905395
5	The Gray House	0.360168408
6	Secrets of the Octopus	0.351524712
7	Gotham's Hero	0.351524712
8	Quisling	0.264965299
9	Archie	0.087949607

Figure 3: Popularity Predictions of Future TV Shows

TIME SERIES GRAPH OF POLITICS SUBREDDITS

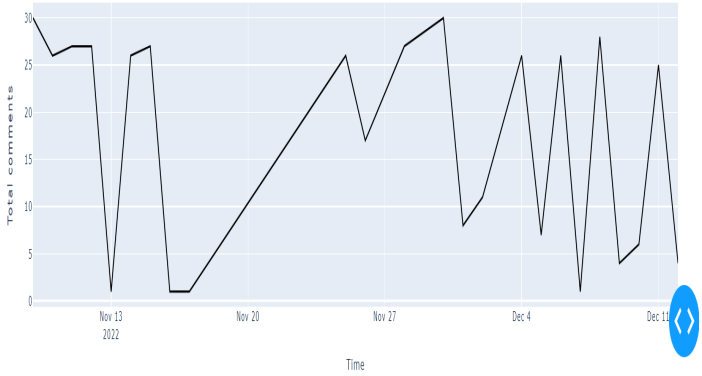

Figure 4: Time Series Graph

6 CONCLUSION

After examining all of these data sets for a range of programs, we were able to predict the factors influencing the popularity of TV shows. And for this project we were able to add some interactivity to basic data characterization by creating a live dashboard.

7 REFERENCES

1. Francis Academic Press, UK – 16, An Analysis of the Factors Affecting the Ratings of TV Series Yan Wang, ISSN 2618-1568 Vol. 2, Issue 4: 16-17, DOI: 10.25236/FAR.2020.020406

2. Andrea Fronzetti Colladon, Maurizio Naldi, 2019, Predicting the performance of TV series through textual and network analysis: The case of Big Bang Theory
https://doi.org/10.1371/journal.pone.0225306

3. TMDB API
https://developers.themoviedb.org/3/getting-started/introduction

4. Reddit Documentation
https://www.reddit.com/dev/api/

5. Twitter Streaming API, resource
https://api.twitter.com/2/tweets/sample/ stream