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 behaviors, news, rising trends, etc. There is a huge amount of user-generated content available, and this alters the environment significantly. The genre, the cast, the budget, the location of production, the number of seasons, the number of episodes, the ratings, and the area 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.

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.

Tmdb is a website that provides millions of movies, TV shows for people to explore. It provides us with the ability to search for the various genre movies, TV shows, the most popular movies or TV shows currently streaming on TV, theaters or available for rent. The platform also provides users to post reviews about the movies/TV shows and provide ratings. The website provides recommendations based on the movies searched or watched by the user. We will make use of the TMDB API which is available on the TMDB API Developers page. The aim of this project is to determine the success of the TV show based on major factors such as 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 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.

2 DATA COLLECTION

This project will be collecting data from 3 data sources, namely, TMDB, Twitter, and Reddit. Below 3 sources specify the details of data collections from 3 sources:

2.1 TMDB

Our data will be gathered from The Movie Database (TMDB), a community-built TV and movie database founded in 2008. Since 2008, the number of submissions to this database has expanded year after year. With over 400,000 developers and businesses using this platform, TMDB has established itself as a top source of metadata.

- GET /tv/tv_id : Would fetch show's budget,cast,region,etc which would help determine their dependency on the show's rank.

- GET /tv/tv_id/changes : Tv shows details by id also provides next and previous episode details.

  GET /tv/tv_id/content_ratings : Individual rating depending on people's liking.

- GET /tv/tv_id/reviews : Individual reviews that would affect the rating of a show.

- GET /tv/top_rated : Would generate the top rated TV shows everyday within our project time period and how they would vary along the way.

- GET /tv/top : API to get most popular TV show list.

- GET /tv/tv_id/watch/providers : Get list of watch providers for a tv show.

2.2 TWITTER

Using Python code, the stream of tweets is collected. Real-time tweets will be gathered using the Twitter Streaming API, resource URL- https://api.twitter.com/2/tweets/sample/stream. A number of parameters that can be helpful for our project's subsequent research will be added to the end of the URL. We will be using the twitter fields available which are: Id - Unique identifier of the Tweet. Tweet - The content of the Tweet. CreatedAt - Creation time of the Tweet Public Metrics - Engagement metrics for the Tweet at the time of the request such as count of retweets, replies and likes. Entities - Contains details about text that has a special meaning in a Tweet such as mentions, hashtags and other details. Annotations - Contains context annotations such as domain,domain name for the Tweet. Language - Language of the Tweet, if detected by Twitter.

The implementation of the collection program used Python's request library for sending the stream request. We have used the Bearer token as a credential. This credential is generated once we have created a developer account on Twitter. The response request is sent using the GET method and we collect the JSON response. We will be collecting every field required for each line in response.

Context annotations are delivered as a context annotations field in the payload. These annotations are inferred based on semantic analysis (keywords, hashtags, handles, etc) of the Tweet text and result in domain and/or entity labels. Context annotations can yield one or many domains.We will use the Domain "TV Show" provided by the Twitter API.

2.3 REDDIT

The task of collecting subreddits will be accomplished by constructing a Python script, but before that, we will be obtaining a temporary OAuth token from Reddit. Then, we must create an API to read the data and stream new postings. Various sub-reddits will be supplied as parameters in order to read or stream the most recent post, generating real-time data. Additionally, a complete dataframe that includes the post's title, date and time, content, and user reactions can be created. A few other parameters, like the following, could be added to further optimize the data:

Limit : in which maximum items that needs to be returned will be given before/after : specify that we only want to search for the posts that are before or after another.

3 NAPKIN MATH

Currently we have data for 4728 TV shows in TMDB, and for approximately 1000 different TV shows' data changes every day, which includes changes in their reviews, TV show details, number of episodes, etc.

4 METHODOLOGY

Following are the steps for data extraction methodology that we are planning to implement:

(1) First, we will generate an API token in the The Movie Database website by creating an account.
(2) Then, we will import all the required packages in Python and initialize the token and key variables. OAuth (Open Authorization) will be used to give a limited access authentication token for authorization to additional resources, for another website or service.
(3) We will be extracting the data based on different API endpoints provided, related to "Budget", "Cast", "Region", "seasons", "reviews", "ratings"","changes"","genre", "people", "popular ","watch providers, etc. The important endpoints that we will use have been mentioned above.
(4) The Data will also be extracted based on certain keywords associated with the TV Shows.
(5) Data processing and cleaning will be done on all the extracted data. Blank or alphanumeric values, comments from spam accounts, stop words, special characters will be removed from the data. Stemming and lemmatization will be performed on the cleaned data to make it efficient to analyze using a machine learning model in order to receive accurate results.
(6) We will store the final data in a NoSQL database.
(7) We will further use this data to analyze, impact of different factors such as cast, budget, region, episode duration, episode count and season count on a movie's success rate as well as popularity.

5 CONCLUSION

The objective of this project is to collect data from 3 data sources-TMDB, Twitter, and Reddit. Data collection is the first and foremost step in building any social media data science pipeline as further analysis is done on the extracted data. Our future scope is to analyze the collected data based on further extraction of primary fields from the raw data, and determine the essential factors involved in contributing towards the success rate of any TV show.