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.

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 thought about and have answered 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

We shall describe our dataset in this part. TMDB, Twitter, and Reddit are the datasets that have been considered, and we will first

provide a brief summary of them. Followed by the approach that have been used to gather 4728 TV shows and over 1000 distinct TV series changes every day, including changes to their ratings, show information, number of episodes, etc in TMDB will then be discussed.

3.1 TMDB API

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 /discover/tv: Discover Tv shows by different types of data like average rating, number of votes, genres, etc.
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.

3.2 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.

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. We have used the following Twitter fields: Id (the unique identifier of the Tweet), Tweet (the content of the Tweet), CreatedAt (the moment the Tweet was created), Entities (Contains information about text that has a special significance in a Tweet such as mentions, hashtags, and other features), Public Metrics (Engagement metrics for the Tweet at the moment of the request such as count of retweets, replies, and likes), Annotations (contains context-specific annotations like the tweet's domain name and language).

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.
- The accessed data is converted to a pickle file in which python object structures can be serialized, further which is the term for the process of transforming an object from memory into a byte stream that can be saved as a binary file on disk. This binary file can be deserialized back into a Python object when it is loaded into a Python program. And further a data frame is formed with the name of df_tmdb from the pickle file.
- 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. Data preprocessing is a Data Mining technique that comprises putting unstructured data into an understandable format. Real-world data is frequently incomplete, inconsistent, lacking in certain behaviors or trends, and contains a great deal of errors. This could lead to the acquisition of poor-quality data, which would then lead to poor-quality models built using the data. Preprocessing data is one way to deal with these issues.
- For Data Preprocessing, firstly Data Cleaning was done in which null, missing and duplicates values were eliminated and One Hot Encoding Technique was used for the categorical data to improve the data accuracy. One-hot encoding in machine learning involves transforming categorical data into a format that can be used by machine learning algorithms.
- After finding unique values for the following columns: genre, type, status, production_companies one-hot encoding was used to make the data appropriate for analysis and to provide correct accuracy.
- 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. This training dataset has both the right inputs and outputs, enabling the model to develop over time. Further, the technique that was used for modelling under supervised learning is "Linear Regression" which is used to find target value based on independent factors.
- 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.
- Further, twitter data was imported for TV shows from the 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 which includes cleaning, tokenizing, and parsing of URLs, Hashtags, Mentions, Reserved words (RT, FAV), Emojis and Smileys.
- Popular TV Shows were fetched using internal API of TMDB i.e., https://api.themoviedb.org/3/.
- Further, top 100 TV shows were fetched to determine their hype on twitter and number of tweets for a mentioned TV show was counted followed by its frequency.
- Based on the tweet frequency top 20 TV Shows were considered to calculate polarity of tweets for a particular show.
- For each of the selected TV show polarity score (positive, negative, or neutral) was calculated based on the tweets and hype was determined based on the polarity score.
- 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.
- Also, analysis for the number of comments for the top 5 subreddit.
- Number of post from a particular domain was also considered.
- 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 DESCRIPTIVE ANALYSIS

Following are the results for the descriptive analysis performed:

5.1 TMDB

Shown below are the results of the analysis performed on TMDB Dataset:

1. To find maximum seasons in a show: 131.0
2. To find minimum seasons in a show: 0.0
3. To find maximum episodes in a show: 15704.0
4. To find minimum episodes in a show: 0.0
5. To find most popular TV Show: Baku Ane: Otouto Shibocchau zo! The Animation
6. To find least popular TV Show: Our Universe
7. Top 20 genres in TMDB for TV Shows are: 'Comedy', 'Documentary', 'Animation', 'Drama', 'Mystery', 'Crime', 'Action Adventure', 'Sci-Fi Fantasy', 'Reality', 'Talk', 'News', 'Kids', 'Family', 'Soap', 'Western', 'War Politics', 'History', 'Music', 'Romance', 'Erotic'.
8. Unique status types in TMDB Database: 'Planned', 'In Production', 'Returning Series', 'Ended', 'Canceled', 'Pilot'.
9. Counted the number of TV shows that are returning Series: 11875
10. Unique TV types present in Dataset: 'Scripted', 'Miniseries', 'Documentary', 'Talk Show', 'Reality', 'News', 'Video'
11. Top 20 Production Companies: 'NBC', 'BBC', 'Warner Bros. Television', 'Universal Television', 'Studio Dragon', 'CBS Studios', 'Amazon Studios', 'TVB', 'Estúdios Globo', 'TV Tokyo', 'NHK', 'Madhouse', 'Toei Company', 'Sony Pictures Television Studios', 'Toei Animation', 'ARTE', 'TV Asahi', 'Aniplex', 'Fuji Television Network', 'SUNRISE'

5.2 Twitter

Shown below are the results of the analysis performed on Twitter Dataset:

The analysis shows that the total number of tweets made are **1064419** and number of unique users who posted the above tweets are **531789**

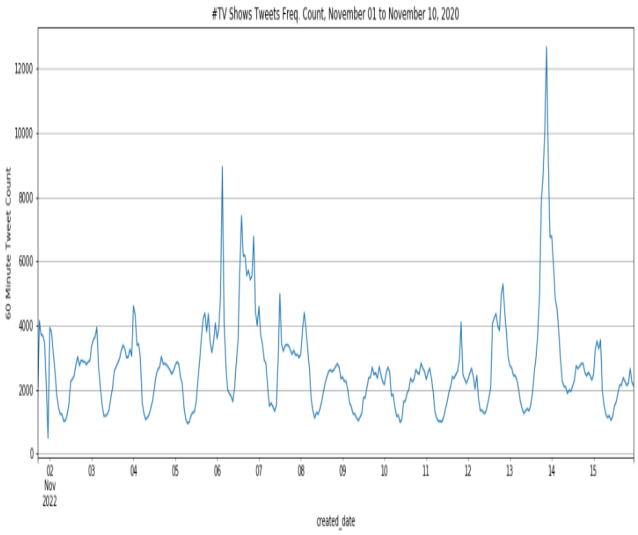

Figure 1: Time Series Plot

In Figure 1: A Time series Graph is plotted to show the number of Tweets received per hour from the 1% stream from 2022-11-01 to 2022-11-15.

5.3 Reddit

Shown below are the results of the analysis performed on Reddit Dataset:

Figure 2 shows the description of the upvote column for the politics sub reddit.

The below **Figure 3** shows the number of comments made per subreddit.

Figure 4 reflect the number of posts made for a particular domain in a sub reddit.

	upvote_ratio
count	1156.000000
mean	0.913512
std	0.102523
min	0.200000
25%	0.890000
50%	0.950000
75%	0.980000
max	1.000000

Figure 2: Up vote Ratio

	number_of_comments
BigBrother	44
thewalkingdead	34
HouseOfCards	23
brooklynninenine	21
television	21
riverdale	20
HIMYM	19
marvelstudios	17
politics	17
blackmirror	14
breakingbad	12
Sherlock	12
gameofthrones	10
30ROCK	8
Letterkenny	5
StrangerThings	4
betterCallSaul	4

Figure 3: Number of Comments

	no_of_post_from_domain
i.redd.it	371
reddit.com	81
v.redd.it	54
self.BigBrother	43
self.thewalkingdead	31
self.HouseOfCards	25
self.riverdale	23
self.HIMYM	23
self.brooklynninenine	22
i.imgur.com	22

Figure 4: No of post from domain

6 DATA VISUALISATION

Below are the plots modelled after the data analysis was performed on the 3 datasets:
Finding hype on twitter for the 100 popular TV Shows:

The below **Figure 5** plots the graph to show the positive and negative hype of a TV show with respect to its polarity. From the plotted graph it can be concluded that **Lisa TV Show** has the highest positive hype and **Breaking Bad TV Show** has the lowest positive hype. Considering the negative hype **Breaking Bad** and the **Walking Dead** takes the same place in the race for the highest negative hype.

Figure 6 above shows a bar graph between various subreddits and no of post made in each of those subreddits. Television and 3D Rock Subreddit were having the highest number of posts.

The line plot below **(Figure 7)** shows top 5 subreddits and the number of comments made for each of the subreddit. **Big Brother** subreddit is having the highest number of comments with **Television** having the minimum.

The **Figure 9** plots the graph for the number of comments made for the politics subreddit from 2022-11-09 to 2022-11-18.

The following 2 figures i.e., **Figure 10 and 11** checks the feature impact on the target wrt the trained model.

Figure 12 shows a graph with respect to prediction and actual test data set.

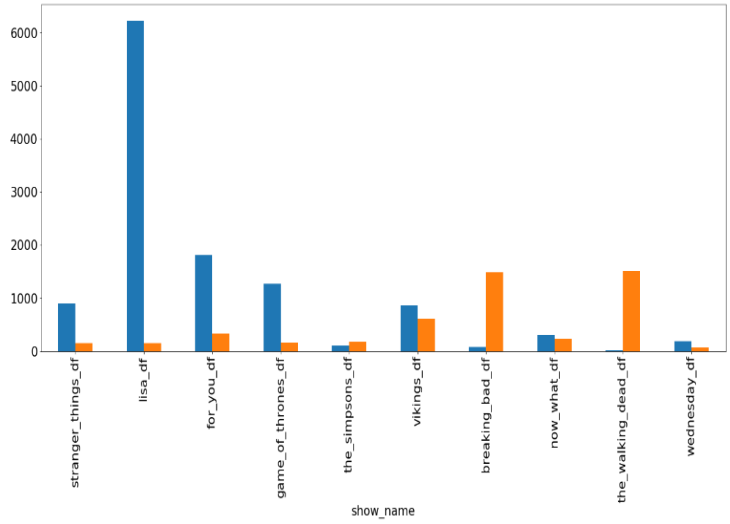

Figure 5: Hype on Twitter

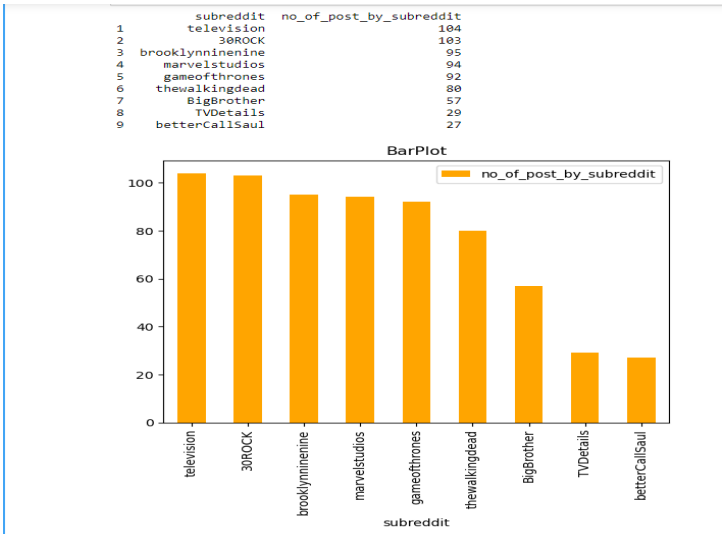

Figure 6: Subreddit vs Number of posts by subreddit

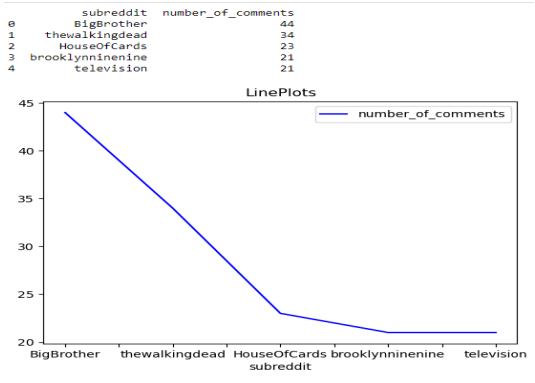

Figure 7: Subreddit vs Number of comments

	_id	dataset	subreddit	created_date	submissions
0	636afd21a7989f77f430dbf0	reddit	politics	2022-11-09	[{'timestamp': 2022-11-09 01:06:41.452000, 'co...
1	636c4091a7989f77f430e192	reddit	politics	2022-11-10	[{'timestamp': 2022-11-10 00:06:41.375000, 'co...
2	636d9211a7989f77f430e6d3	reddit	politics	2022-11-11	[{'timestamp': 2022-11-11 00:06:41.442000, 'co...
3	636ee391a7989f77f430ed72	reddit	politics	2022-11-12	[{'timestamp': 2022-11-12 00:06:41.366000, 'co...
4	63703511a7989f77f430f37a	reddit	politics	2022-11-13	[{'timestamp': 2022-11-13 00:06:41.412000, 'co...
5	63718691a7989f77f430f8ee	reddit	politics	2022-11-14	[{'timestamp': 2022-11-14 00:06:41.355000, 'co...
6	6372d811a7989f77f430feb2	reddit	politics	2022-11-15	[{'timestamp': 2022-11-15 00:06:41.332000, 'co...
7	63742991a7989f77f4310441	reddit	politics	2022-11-16	[{'timestamp': 2022-11-16 00:06:41.348000, 'co...
8	63757b11a7989f77f4310933	reddit	politics	2022-11-17	[{'timestamp': 2022-11-17 00:06:41.691000, 'co...
9	63810b89b3f5bca534a6d4c4	reddit	politics	2022-11-25	[{'timestamp': 2022-11-25 18:38:01.124000, 'co...
10	63815fe9b3f5bca534a6d561	reddit	politics	2022-11-26	[{'timestamp': 2022-11-26 00:38:01.109000, 'co...
11	6382b169b3f5bca534a6dc44	reddit	politics	2022-11-27	[{'timestamp': 2022-11-27 00:38:01.115000, 'co...

Figure 8: Dataframe for the politics subreddit analysis

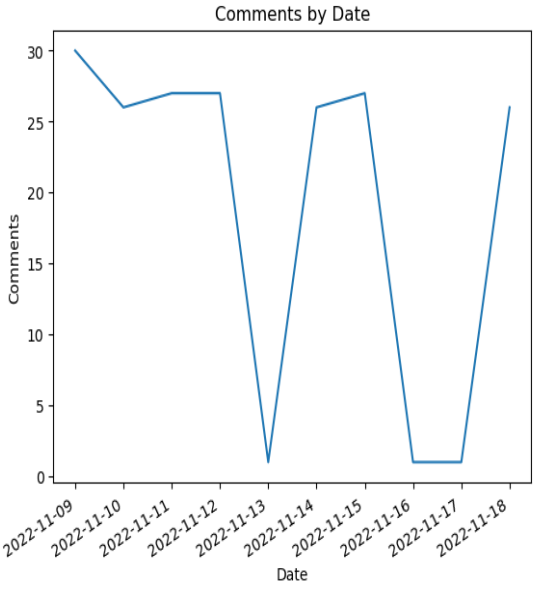

Figure 9: Comments by Date

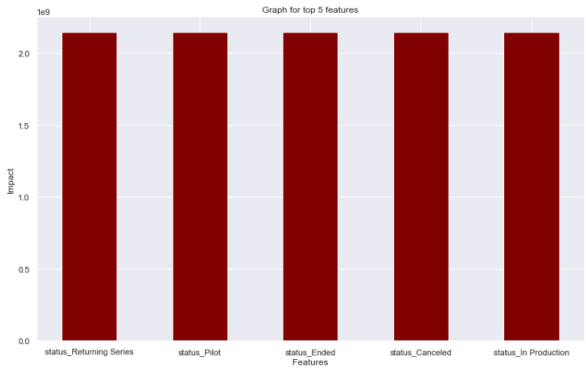

Figure 10: Bar Graph of Top 5 features vs Impact

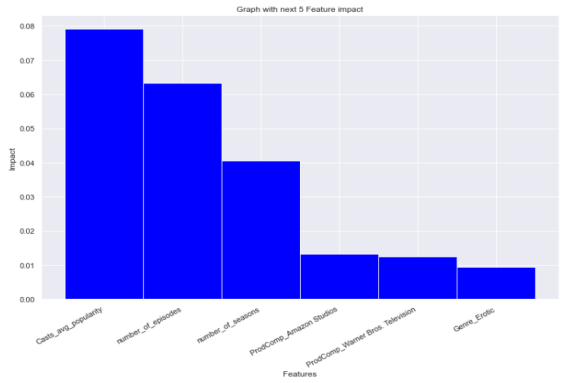

Figure 11: Bar Graph of Next 5 features vs Impact

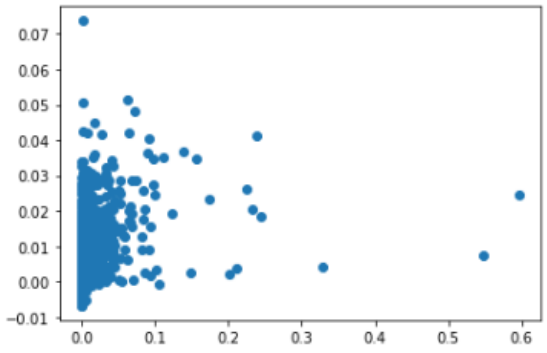

Figure 12: Graph on predictions and actual test set data

Figure 13 reflects a histogram on validation dataset and predictions made.

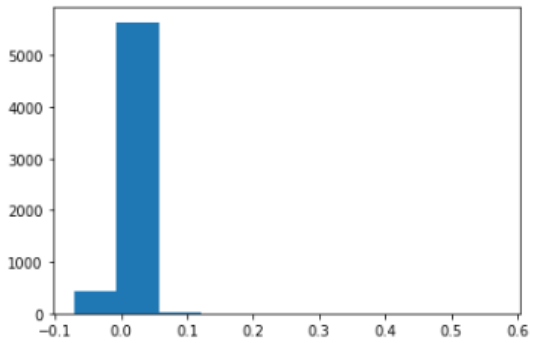

Figure 13: Histogram for TMDB Data

The below **Figure 14** shows the number of comments made per subreddit.

Figure 15 shows the plotting of the PDF and CDF for the number of comments made for each of the subreddit (Reference Figure 14).

Figure 16 shows the word cloud for a subreddit.

	number_of_comments
BigBrother	44
thewalkingdead	34
HouseOfCards	23
brooklynninenine	21
television	21
riverdale	20
HIMYM	19
marvelstudios	17
politics	17
blackmirror	14
breakingbad	12
Sherlock	12
gameofthrones	10
30ROCK	8
Letterkenny	5
StrangerThings	4
betterCallSaul	4

Figure 14: Number of Comments

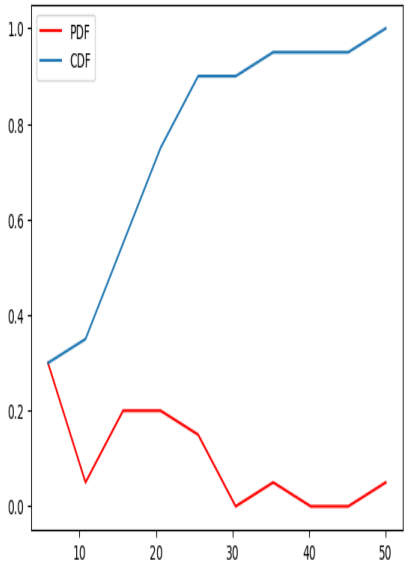

Figure 15: PDF CDF for the number of comments

Figure 16: Wordcloud for Stranger Things TV Show Subreddit

7 CONCLUSION

In summary, we find that:

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.

8 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