Visualizing Different Biases Between Television Series, Movies along with the Sentiment Analysis of Different Reviews

Ganesh More
Student of Masters in Data Analytics
National College of Ireland
Dublin, Ireland
x18180451@student.ncirl.ie

Ameya Dalal
Student of Masters in Data Analytics
National College of Ireland
Dublin, Ireland
x18184430@student.ncirl.ie

Shirish Sonvane
Student of Masters in Data Analytics
National College of Ireland
Dublin, Ireland
x19164165@student.ncirl.ie

***Abstract*—Analyzing various factors and trends between TV shows and movies is the goal of this paper with the sentiment analysis of their reviews. Visualizing is the key way to interpret and analyze the data. With the help of visualization, it is trouble-free to discover patterns and trends present within the data. By taking, the count of movies and TV shows released throughout the years along with their genres, runtime, fame, reviews, and different other essential factors into consideration, visualizations have been formed to understand more clearly what are the present trends in the world. Alongside, a sentimental analysis of the different reviews of movies and TV shows has been conducted, which later distinguished into positive and negative in terms of words. From this sentiment analysis, it could be said that the movies have almost equal aspects that are positive and negative. This same analysis is applicable for TV shows as well. In terms of the analysis, count of the movies and TV shows aired in recent years is big compared to the earlier years. Drama and Comedy are the most trendy genres among the people of movies and shows respectively.**

***Index Terms*—Visualization, Ratings, Vote Count, Genres, TV Shows, Entertainment, Genres, Movies.**

I. Introduction

Movies, as well as tv shows, are nowadays a very common way of keeping oneself entertained in these hectic days. It is the best way to relax your mind from all busy lives. Everyone around the world has some movies or shows which they like or some which they do not. It is always fascinating to see that there are some specific tv shows and movies that are considered and most watched by the individuals.

Well, it is very obvious to pop a question that why there is this pattern? There has to be some factors that affect one's attention in watching only particular shows or movies. To understand this in detail, some data of movies and shows were collected, released throughout the years. To study the patterns and popularity, some graphs are being plotted using various aspects between them.

One more analysis was taken into consideration, to understand what peers think about the movies and shows. This analysis conducted on the reviews of these movies and shows released. These reviews of each movie and shows contain some good and bad aspects about them. While performing an analysis on the three datasets, we came up with a below research questions to which we tried to find answers. " Which genres are more popular among tv shows and movies? Which language is more popular between tv shows or movies? "

To get an understanding of what types of tv shows and movies are famous among the public, we need to answer these questions. Theses answers might give some perceptions about them. For descriptive analysis, different attributes like production companies and countries, title, runtime, genres were being taken into thought. The reviews data divided into positive reviews and negative data to analyze the same, so that we can retrieve some meaningful insights. The next section describes the methodology implemented in this research and understandings to answer the above questions.

II. Methodology

We have used python as the programming language. It is a general-purpose and open-source. Python has a strong and large number of libraries. Python is compatible with almost all platforms and operating systems. Python has the ability of garbage collection in order, to free memory allocated to objects which are no longer in use. Python also has abilities like text mining, handling a large number of records. We have extracted the data from TMDB (The Movie Database) using its API. TMDB provides various APIs to fetch the data like movies, tv shows, ratings, people etc. TMDB has movies and tv shows data from around the world. Before fetching the data from TMDB, we created an account on it and requested the key.

For the authentication so that we can communicate with their server. We have fetched only tv shows and movie data for this project. Since the data that we fetched was in JSON format, we could easily store the same in the MongoDB. JSON format is like the dictionary in the python and MongoDB supports documents database which means it stores the information in JSON format. Storing data in this format is a more powerful and expressive way to present the data rather than row and column format. Since we wanted the 1000 show and movie IDs, we used the loop starting from

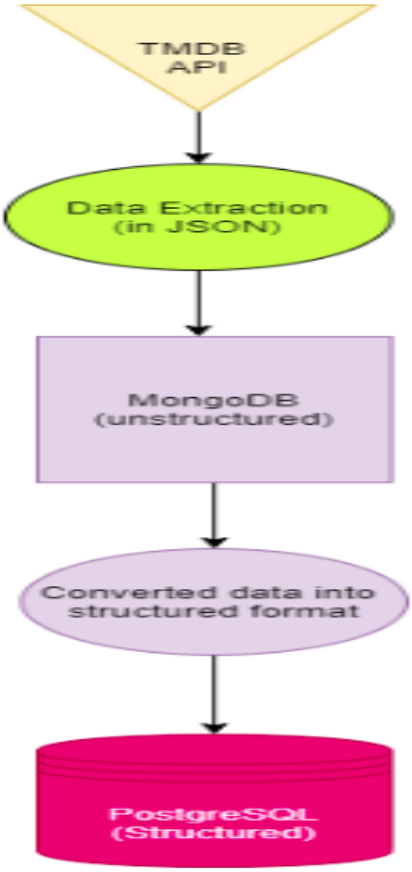

Fig. 1. Flow of Data

1 till we get the 1000 rows of the data from TMDB. The data which was being fetched was in the unstructured format. Since MongoDB is NO SQL database it supports the unstructured data to store in. Also, it is an open-source for the storage of dynamic information in the format of the JSON document. We created four collections for three datasets under the database 'dap_project'.

- dap_movies
- dap_movies_review
- dap_tv_shows
- dap_tv_reviews

Before doing so, we created an ec2 instance on AWS using Debian 10 buster AMI. We have used AWS to store the data. The reason behind this is that anyone from the group can fetch the data from the database to operate on it. We installed MongoDB and PostgreSQL on that instance and connected to them using python. After this we created the database named 'dap_project' in the PostgreSQL to store the structured data in it. Postgres is also open-source and RDMS. It safely stores the most complicated data in tabular format so that one can retrieve the same just using SQL queries as per their need. Before transforming the data into tabular format we removed some symbols and unwanted strings from it using the regex

Attributes of movies
id
genres
production companies
production countries
release date
runtime
languages
title
vote average
vote count

Attributes of tv shows
id
genres
production companies
vote average
runtime
languages

Attributes of movie reviews
id
author
content
review id
url

Attributes of tv reviews
Id
Author
Content
Review id
url

Fig. 2. Attributes of the table.

from unstructured data. We created the four more tables in PostgreSQL as well named below.

- movies_reviews
- movies
- tv_shows
- tv_reviews

We used SQL queries each time to fetch the data from the database while plotting the graphs. Clauses like group by, order by, and functions like count, average has been used here. For the sentiment analysis of the reviews NLTK package has been used. Using the libraries present in this package we have separated the positive and negative data.

III. Related work

One of the important packages used for sentiment analysis is NLTK (Natural Language Toolkit). It is a platform that is used for constructing a python program for implementing in Natural Language Processing (NLP) [1]. Tokenization, parsing, classification, stemming is some of the libraries of NLP used during Sentiment Analysis. It helps while processing the review for a particular object and it is fetched in the form of a label. Another package used by Goyal and A. Parulekar, is very useful while predicting the review of movies and TV shows, classification of the reviews is also carried out in positive, negative, and neutral [2]. Likewise, the word cloud was implemented in this code for displaying positive and negative words.

Nowadays we can reach people regarding reviews through social media. For reducing the rating of other movies, some people tamper the reviews which can be observed online [3]. A conclusion can be made by observing the review and the rating of movies and TV shows which plays an important role in generating the revenue of the movie. Rating and the reviews of the movies from the data with the description are useful for analyzing and for visualization. Commendation system for TV shows and movies is an important feature depending on reviews. It helps in featuring and making a short story about the movie i.e. good words and bad words.

Database stored wealth information, its vast amount of to explore manually by browsing for Example. In September

2019, TMDB and IMDB contained 10 million movie titles 7 million TV shows, and information about 50 million registered users. To analysis, the data using visualization tool.to collecting the in amount and provide for the user, some web portal is beneficial just like IMDB and TMDB. The most crucial factor is to analyze the data and review these web portals are essential parts for the production companies. Visualizing the data from the movie and tv shows database.it involves the combination of several things. The visual concept gives a way to represent the data in many views [4].

Weimao Ke and Bruce herr they show the relationship between movies and major co-actors [5]. They visualized the relationship using plots.so others can understand the visualization is a significant part. They divide Genre into seven different color types to imagine in different colors, for visualization have different graphs. They have mainly used matplot and geoplotlib libraries. Using these two libraries can plot different kinds of figures, .user can show differently or represent their view in front of others[]. to keep visualization

IV. Understanding using visualization

Let's start with the basic visualization, firstly we visualized total no. of movies and Tv shows in each Genre.

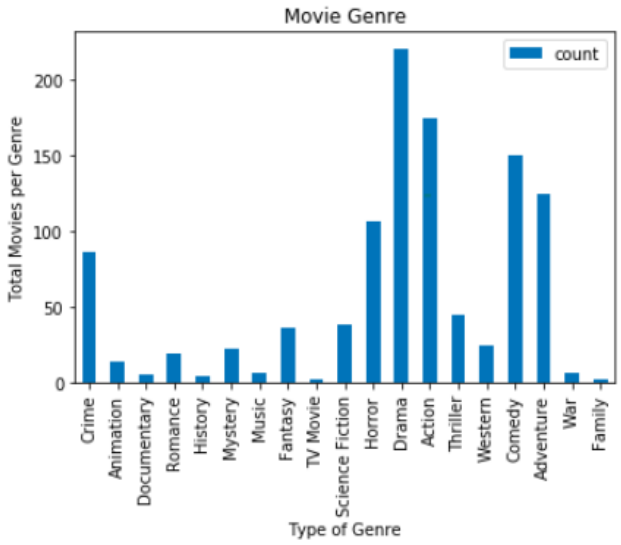

Fig. 3. TV Shows vs Genre

From both graphs, we noticed in the Movie industries Drama movies are having more popularity, and family movies have low demand. Instead of Drama, people also like to watch Action, Comedy, and adventure movies. The Drama films have 200+ movies.

In the TV shows people most liked to the comedy shows and rather than Drama, animated, mixed combination of action and adventure also popular tv shows. But the only action shows they won't want to watch.

Television industries are having more shows because nowadays we have lots of media to watch instead of going to the theater and watch movies.

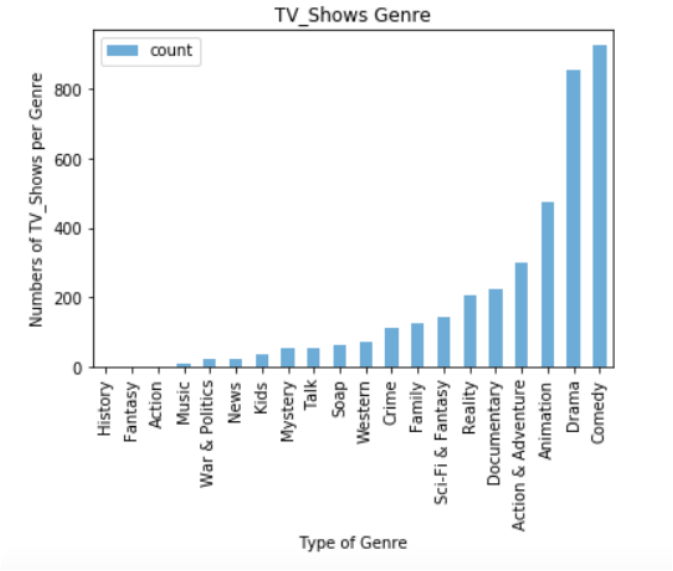

Fig. 4. TV Shows vs Genre

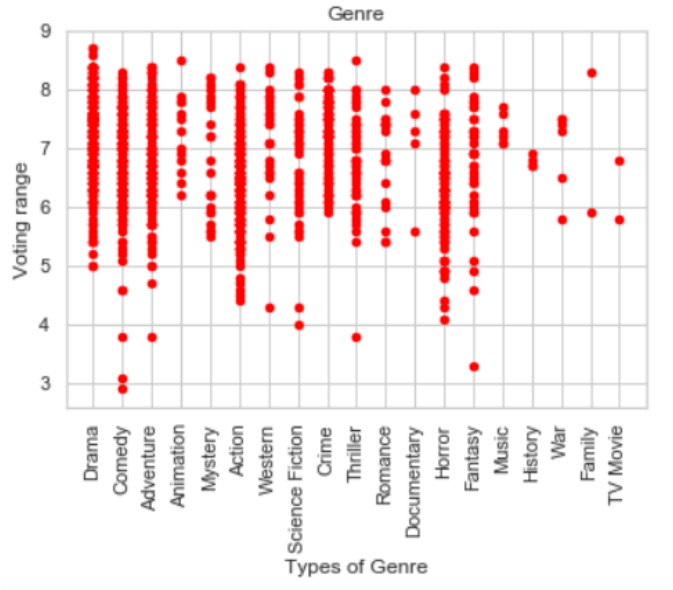

Fig. 5. Movie Genres

In fig.6 drama movies have most voted movie types. The average votes on film were between 5 to 8. Comedy, Adventure, Action & Horror movies had a most anticipated film. But if we noticed TV movies having fewer votes maybe instead of watching movies on television, they preferred to watch in a theater.

Animation, Comedy, Drama, Reality, Talk, Action & Adventures TV shows had all range of votes. Kids also enjoying there their television shows. It looks like some of these TV shows are not so popular like Action, Music, History, and Fantasy. These TV shows are popular and not so popular. Because they had a smaller number of votes.

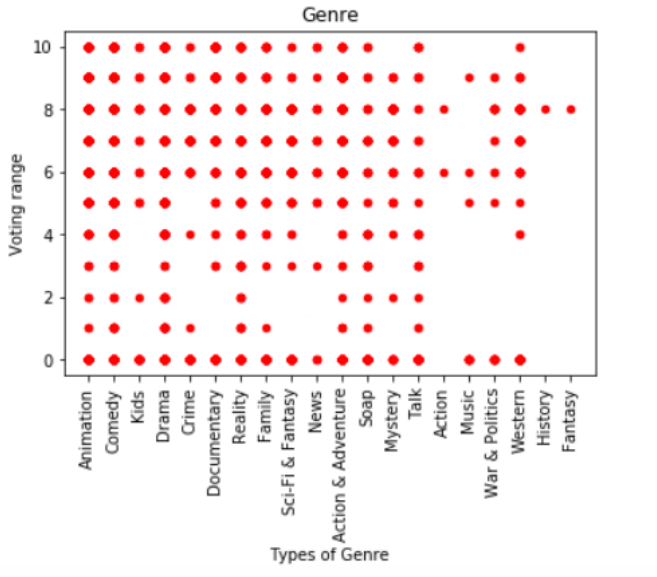

Fig. 6. TV Show Genres

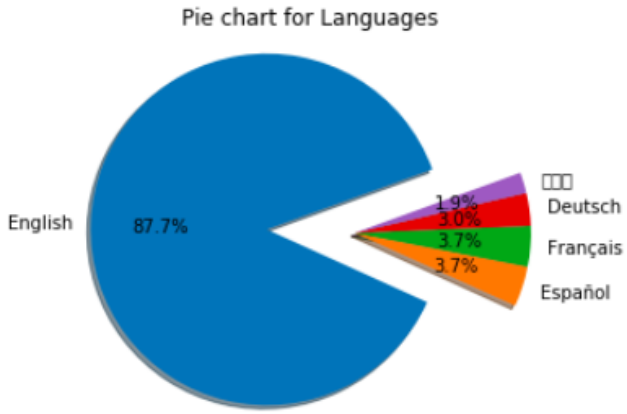

Fig. 7. Top Seven Languages of Movies

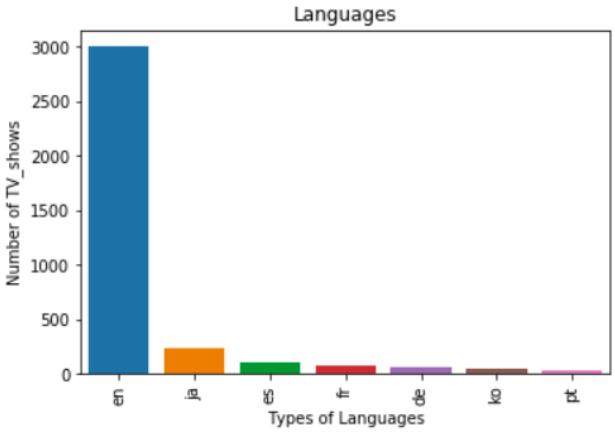

Fig. 8. Top Seven Languages of TV Shows

In movie industries, they produce all language movies. But most of the film in the English language. We can see in the movie industries that production companies produced 88% of their movies in English. So, we can say that English movies are more popular than in other languages. Española, Francia's, and Deutsch having a cover the audience and remains coved by their regional language.

In the Tv industries, the English language has most tv shows, and Portuguese languages are having a lesser number of tv shows. If we see both FIGURES NAME, we found the English language had the most popularity in movies and also in tv shows. As we can see, 88% (1082) in films and approx. 3000 tv shows are in English.

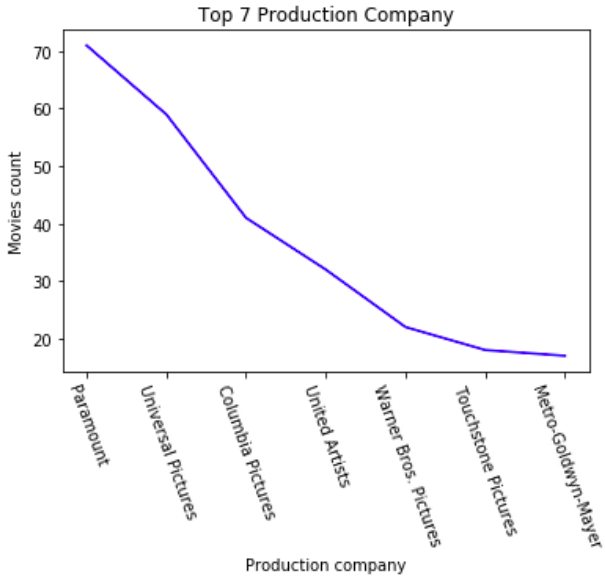

Fig. 9. Top Seven Production Companies in Movie Industry

From both images, we noticed that in the top 7 production company, the Paramount production company produced a higher number of movies, and Metro-Goldwyn-Mayer company had lower no. Of production.in the movie industries Paramount is more popular than others. It produced almost 70 films.

On the other hand, TV industries have a BBC company that produced above 120 TV shows. Universal Television is a tough competitor for BBC.these two are having more popularity instead of this Televisa made below 40 TV shows.BBC is a giant company for Televisa.

In the above graph, we took the top 10 movie production countries the United States of America has the most movies, and New Zealand has the lowest movies. Instead of the USA United Kingdom is the second most movie production country. Rather than Australia, Japan, Canada has the same product range.

In this graph, we tried to show how much movie's release in a particular year? The movie industry keeps increasing from 1991 to 2006. In 1991 the above 15 movies were released, and

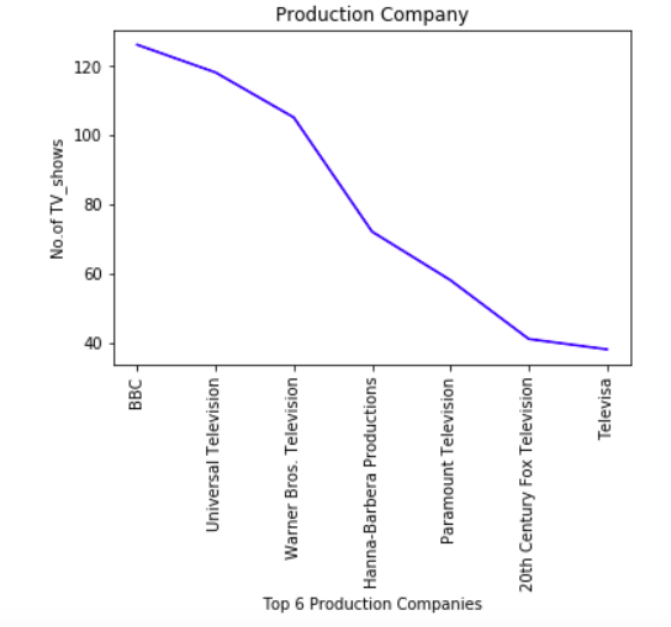

Fig. 10. Top Seven Production Companies in TV Industry

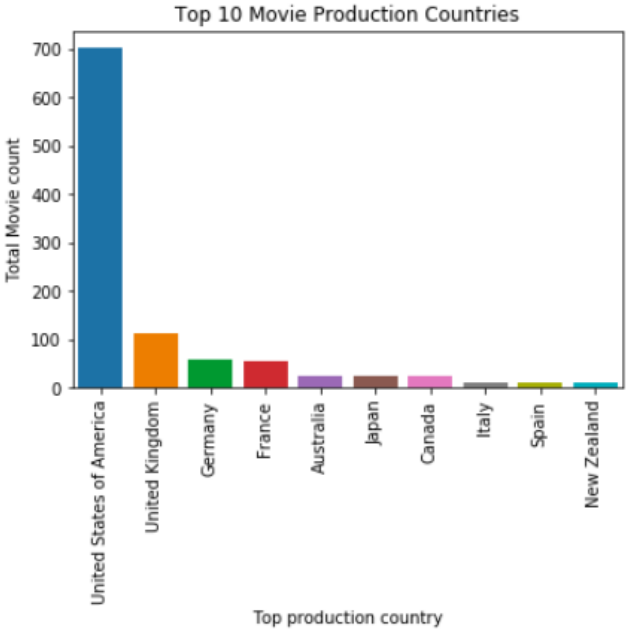

Fig. 11. Top Ten Production Countries in TV Industry

in 2006 the rate of the film released almost double. It shows how movie industries are growing year by year.

This bar graph shows the top 8 most voted movies of all time on TMDB. On the X-axis, we have movies title, and, on the Y-axis, we have a vote count per film. From this graph, we can observe and state that 'The Dark Knight' is the most voted movie, and 'The Matrix' is least voted out of the top 8 most voted movies. Harry Potter and the Philosopher's Stone, The Lord of the Ring, and Titanic these films got almost the

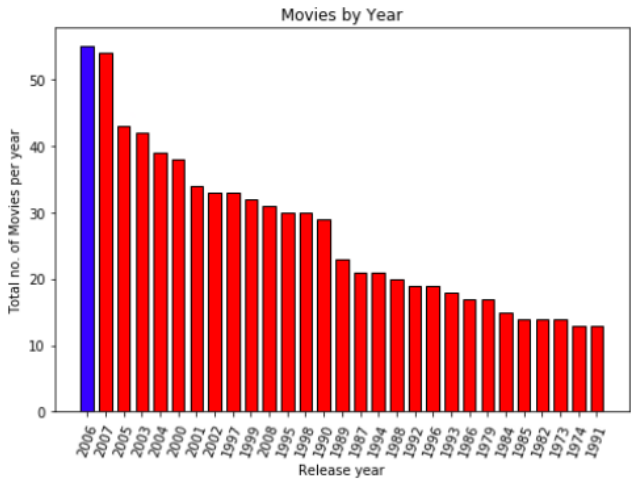

Fig. 12. No. of Movies Released per Year

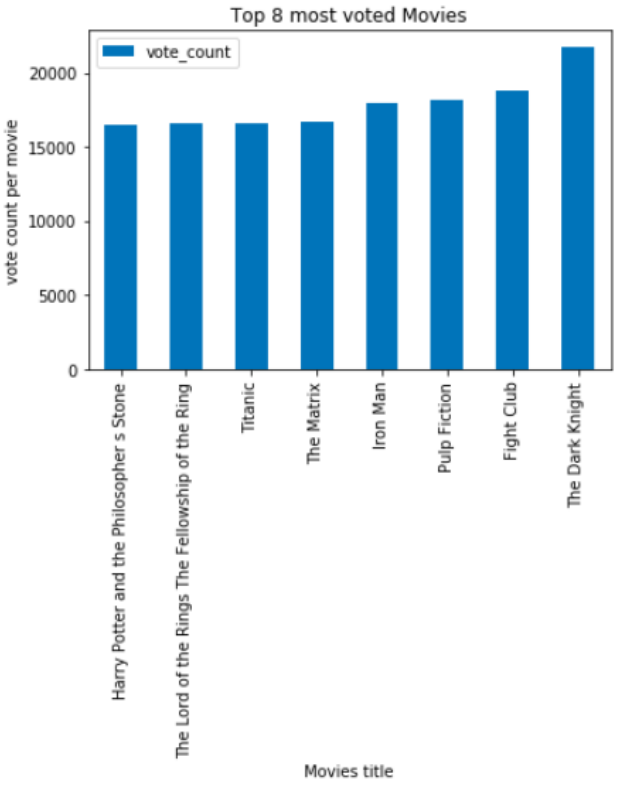

Fig. 13. Top Eight Most Voted Movies

same votes.

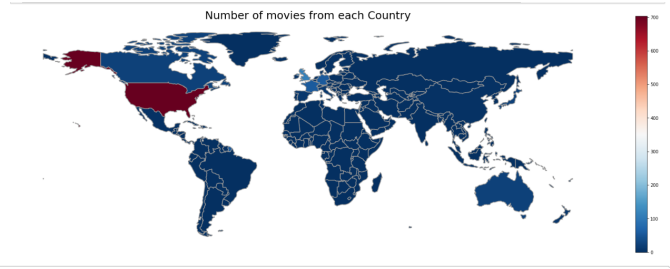

Fig. 14. Movies per Country on World Map

Fig. 15. Positive words

This is the world map of movies. It shows that no. of movies release per country. The United States of America and the United Kingdom have released the highest no. of movies.

Sentiment analysis is used for understanding the way of thinking of people and their views or attitude that is communicated in text. In this project, we studied the technique of sentiment analysis by using some method; sentiment analysis can be done if the data is in the text format. Inner join can be can used for sentiment analysis while fetching the data by using SQL queries. The positive words and the negative words are shown in the python notebook. While doing the sentiment analysis of both movies and TV shows we can observe that a greater number of negative words were noticed as compared with positive words. During the analysis we can see that the popularity of the data is dependent on the reviews.

Wordcloud is a method that is used is doing a presentation of data visually. Word cloud is used for spotting the word frequencies which are popularly used. The word which is more frequently used are displayed in large and bold size. Considering the word cloud, it shows the positive words with high frequency that were given by the people. We can conclude from this that the review supports an important role in the popularity of tv shows and movies. The count of the word indicated that people do have more interest in watching movies than tv shows.

Fig. 16. Negative Words

The next graph shows the word count with frequencies of negative words used for movies used in the review. While plotting we used image format to specify the shape of the words by uploading the image file into python.

Fig. 17. TV Positive Words

In the next graph, let us see the visualization done on tv shows. Here we will display the positive words which the people used more frequently in reviews.

In this figure, we will display the negative words used for review regarding tv shows.

V. Conclusion

By looking at the review information of review data and sentiment analysis has affected the popularity of the movies and tv shows. Comparison between popular movies and other movies, the popular movie has more positive reviews and negative reviews. The machine learning model can be applied to

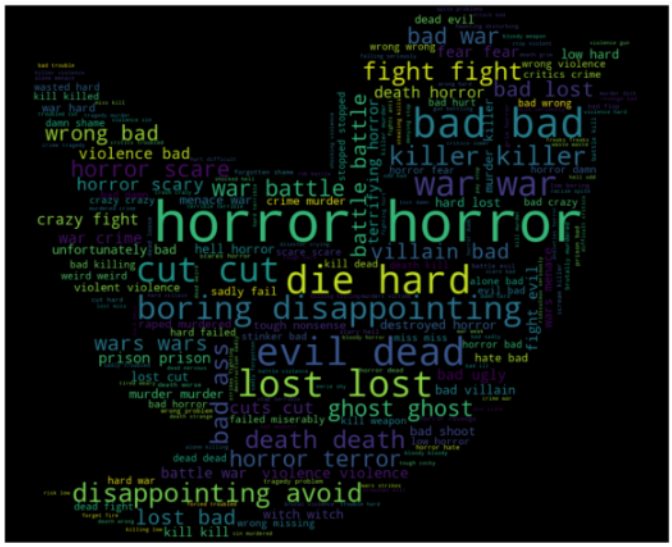

Fig. 18. TV Negative Words

this sentiment analysis since we have split the words into three sections: positive, negative, and neutral. With this information, we can show visualization with other factors that were not included which can give us more insight. Age, gross network, estimated budget, etc. can be included in the visualization for better insight. Machine learning methods can also be used for predicting the popularity of the upcoming movies based on the review statistics about the people's attitude and it can be useful for making changes while realizing the movie or tv show for better results and visualizations.

Considering all the information from the plots, it can be specified that all the variables included show the trend depending over years. As it can be seen that movies related to drama and tv shows related to Comedy genre are the most popular or most viewed within remaining data of movies and tv shows. From the Bar graph, English is most common language used within tv shows and movies in all the countries. From this it can be concluded that the questions which were mentioned earlier in the report were answered through visualizations.

REFERENCES

[1] A. N. Jebaseeli and E. Kirubakaran, "A survey on sentiment analysis of (product) reviews," *International Journal of Computer Applications*, vol. 47, no. 11, 2012.

[2] A. Goyal and A. Parulekar, "Sentiment analysis for movie reviews," 2015.

[3] V. Parkhe and B. Biswas, "Genre specific aspect based sentiment analysis of movie reviews," in *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 2418–2422, IEEE, 2015.

[4] J. Lee, G. Noh, and C.-k. Kim, "Analysis & visualization on movie's popularity and reviews," in *2014 International Conference on Big Data and Smart Computing (BIGCOMP)*, pp. 189–190, IEEE, 2014.

[5] B. W. Herr, W. Ke, E. Hardy, and K. Borner, "Movies and actors: Mapping the internet movie database," in *2007 11th International Conference Information Visualization (IV'07)*, pp. 465–469, IEEE, 2007.