Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

# Exploratory and Sentiment Analysis of Netflix Data

Karthik Babu Vadloori[1]
BTech (Computer Science Engineering)
Sreenidhi Institute of Science and Technology[1]
Hyderabad, India – 501301[1]

Shriya Madhavi Sanghishetty[2]
BTech (Computer Science Engineering)
Malla Reddy Institute of Engineering and Technology[2]
Hyderabad, India – 500100[2]

***Abstract*—The term "Exploratory and Sentiment Analysis" is a conjunction of two separately unique approaches present in the vast field of Data Science. The key to this project is to enhance the value of the Data being utilized, in our case it is Netflix Data – which is an Open-Source Data Set obtained from Kaggle – that was wrangled and exercised to derive maximum insights using EDA – Exploratory Data Analysis and Sentiment Analysis after the amalgamation of two additional sets – Geographical Latitudes & Longitudes and Netflix Title Critics/Reviews Data Set. The project is made using different utility analytical tools present in Python Library of versatile packages. This paper introduces systematic and insightful usage of methods for Exploratory Data Analysis & Sentiment Analysis by utilizing various packages concerned.**

***Keywords*—Exploratory Data Analysis; Sentiment Analysis; Data Analytics; Python; Seaborn; Numpy; Tensorflow - Keras**

## I. INTRODUCTION

The term "Data Analysis" is known to be rooted in the statistics space, which itself is known to have a long history. With the help of the statistical development techniques, we can derive interesting outcomes. The advancement of rapid technological implications in the world led to a consequent advent of Big Data; we are constantly being faced with enormous amounts of raw data which is subject to future enhancements based on the required parameters and criteria by an entity. Starting with the collection of data, the most common and subsequent step is to perform the analysis of it. Data analysis is hence known to be a scientific process solely focused on the data as its subject. It begins with retrieving data from various external-cum-internal sources and then performing intrinsic analysis with the data in order to discover and obtain beneficial information catering the needs of an entity. For example, the analysis of population growth by district can help governments determine the number of hospitals that would be needed in a given area. When collecting the optimal data for analysis it must hold the minimum viability in terms of features and attributes suitable for our analysis. This can be represented in terms of bodily and health-oriented features like Health Status, Age, Male:Female Ratio, BMI etc., will provide much more issue specific insights over the population. It can enable a person to visually represent these features as per the requirements. Fundamentally, there are two primary methods for data analysis – based on the nature and characteristic of data - qualitative data analysis and quantitative data analysis techniques. These data analysis techniques have the scope to be utilized independently or in combination with other methods in order to gain access to some of the best business and intelligence-oriented insights for making better decisions over the already present data.

## DATA ANALYSIS AS A SUBJECTIVE MATTER:

### *A. Requirements Gathering*

The data is the necessary requirement for providing inputs to any type of analysis. It can be based on the requirements and parameters based on the user. This data can be either numerical or categorical. The purpose and scope can range from supervised to unsupervised learning.

### *B. Data Collection*

The required data can be collected from a wide range of sources. It can be structured based on the criteria provided by analysts to custodians of a particular data set. The data can be man-made or in the form of technological output over utility (sensor system tracking) and many other implications.

### *C. Data Processing & Cleaning*

The data when purported for utilization must be processed on the level where the needs for analysis are satisfied. This includes placing data in the form of rows and columns that are human-understandable in nature. Further, it must be cleaned for getting rid of any redundant data, or minimalize the presence of anomalies prior to the deployment of the data for analysis. The figure given below explains these four fundamental steps in a lucrative pictogram.

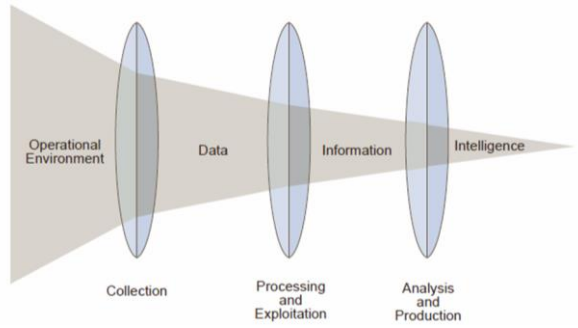



Figure 1.1: Relationship of Data, Information and Intelligence. Source: Wikipedia

## II. LITERATURE SURVEY

With the advent of technology, the need for consumption of data has increased tremendously. Every single activity of our life has become interlinked with data. As the famous British Mathematician – Clive Humby – once said,

"Data is the new oil."

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

As interesting as it may sound, the fact that there is a constant need for intelligent solutions – with data as inputs – cannot be neglected. Due to these changes, we have witnessed a paradigm shift in almost every single field around the world. One such field is the entertainment industry that had a rapid transposition as the industry soon began to adopt virtual methods for releasing its content. The Over-The-Top (OTT) as a means to provide and deliver content has gained huge prominence around the world. One of the key players in the game being Netflix, has changed the way people subject themselves to entertainment. The from-home-all-in-one package type subscription services that these OTT platforms has to offer had customers glued to them quite literally. All these behavioral information about large base of customers is essentially a valuable data. Thus, our project focuses on deriving crucial insights from the publicly available Netflix dataset (obtained from Kaggle[6]). We have also made use of Geographical Data Set and Netflix Title Reviews Data Set.

## III. EXPLORATORY DATA ANALYSIS

Once our required datasets are generated – with inculcated properties like optimal, structured and human-understandable data format – we can now proceed further for an in-depth analysis of it. To begin with, we have chosen EDA as our primary step to analyze our data. We have applied a variety of techniques to gain maximum insights from our data set.

### *A. Correlation Heat Map*

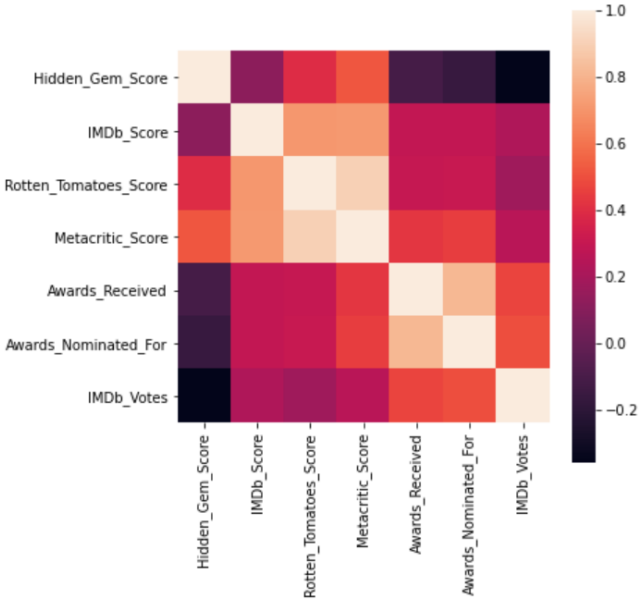



Figure 3.1: Correlation Heat Map of Netflix Data

From figure 3.1, the Darker colored regions represent less correlation and lightest color represents highest correlation. Which essentially means that the above selected attributes are interrelated to such an extent that indicates us about the dependency factor within them. Thus, medium to high ranged correlation attributes can be utilized for our analysis to derive interesting patterns. The features like IMDb_Votes and IMDb_Score have medium red color which is neither too dark nor too light in contrast, indicating us that there is a variation in terms of them both i.e Number of Votes doesn't directly mean IMDb_Score is higher, the quantity of the votes may have either negative or positive effect on the Score. This explains the way all the attributes are correlated.

### *B. SNS Darkgrid Plotting – Title Format*

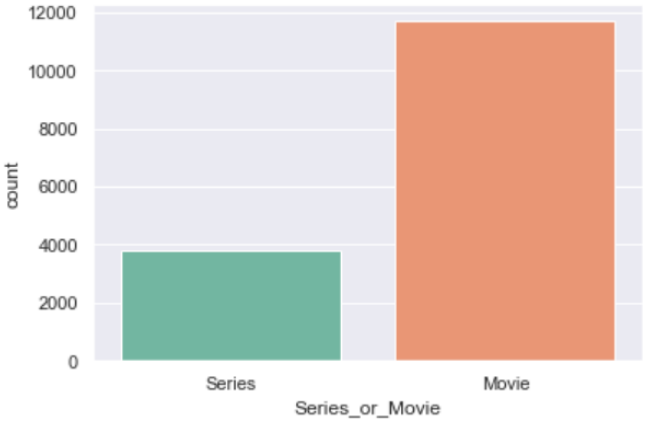



Figure 3.2: SNS Darkgrid Plotting Netflix Data

From the figure 3.2, we can understand the two main categories of formats that are present in Netflix. Using the SNS plotting method we can find out that Number of Movies are about 2.5 times the Number of Series in the Netflix platform. The x-axis has the labels for Series & Movie, where as the y-axis consists of the count of these formats.

### *C. SNS Plot – IMDb Rating*

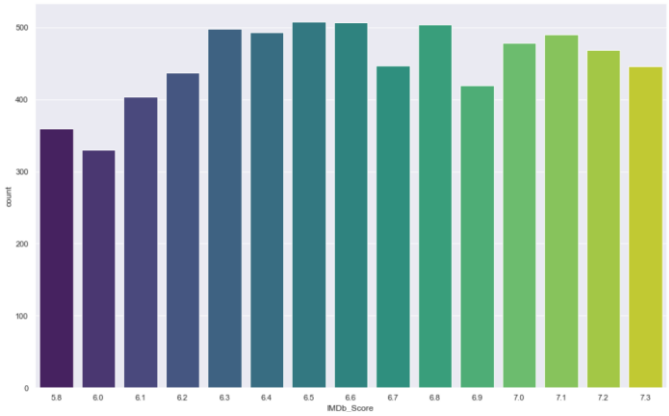



Figure 3.3: SNS Darkgrid Plotting - IMDb Score

From the figure 3.3, we can understand the range of titles – both Series and Movies combined – in terms of their IMDb score. We can comprehend that the highest number of titles range from IMDB_Score 6.4 to 6.8 in the given dataset. The x-axis has the label for IMDb_Score, whereas the y-axis consists of the count.

### *D. Word Cloud Representation*


Figure 3.4.1: Word Cloud Representation – Genres (unmasked)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021


Figure 3.4.2: Word Cloud Representation – Genres (mask)

- Netflix Logo is utilized as the mask.

In the figures 3.4.1 and 3.4.2, we have generated WordCloud for the first case, where the relevance of Genres is mapped (highest count genre is represented in larger font size and lower relevance is smaller). Similarly using "mask=img" feature in the WordCloud method we have mapped our genres into the Netflix logo for better graphical representation.

### *E. Title Categorization – Super Hit, Hit, Average & Flop*

| | Director | SuperHit |
|---|---|---|
| 0 | Akira Kurosawa | Rashomon |
| 1 | Alex Gibney | Going Clear: Scientology and the Prison of Belief |
| 2 | Alfred Hitchcock | Rebecca,Psycho |
| 3 | Amole Gupte, Aamir Khan | Taare Zameen Par |
| 4 | Andrew Lau, Alan Mak | Infernal Affairs |
| 5 | Andrey Zvyagintsev | The Return |
| 6 | Aniruddha Roy Chowdhury | Pink |
| 7 | Anubhav Sinha | Article 15 |
| 8 | Anurag Basu | Barfi! |
| 9 | Anurag Kashyap | Gangs of Wasseypur: Part 2,Gangs of Wasseypur:... |

Figure 3.5.1: SuperHit Titles – df.head(10)

| | Director | Hit |
|---|---|---|
| 4 | Andrew Lau, Alan Mak | Infernal Affairs 3,Infernal Affairs 2 |
| 8 | Anurag Basu | Life in a ... Metro |
| 9 | Anurag Kashyap | Raman Raghav 2.0 |
| 10 | Ashutosh Gowariker | Jodhaa Akbar |
| 13 | Bong Joon Ho | Okja |
| 17 | Bryan Singer | Valkyrie,Superman Returns |
| 18 | Chan-wook Park | Joint Security Area,Thirst,Sympathy for Mr. Ve... |
| 22 | Clint Eastwood | Richard Jewell,The Mule,The Bridges of Madison... |
| 24 | Danny Boyle | Yesterday,T2: Trainspotting,Steve Jobs,Sunshin... |
| 25 | Darren Aronofsky | Mother!,Pi |

Figure 3.5.2: Hit Titles – df.head(10)

| | Director | Average |
|---|---|---|
| 10 | Ashutosh Gowariker | Mohenjo Daro |
| 15 | Brian Taylor | Mom and Dad |
| 22 | Clint Eastwood | The 15:17 to Paris |
| 58 | John G. Avildsen | Rocky V |
| 62 | Jonathan Demme | Ricki and the Flash |
| 64 | José Padilha | 7 Days in Entebbe |
| 74 | Lilly Wachowski, Lana Wachowski | Jupiter Ascending,Speed Racer |
| 80 | Martin Campbell | Green Lantern |
| 85 | Michel Gondry | The Green Hornet |
| 108 | Ridley Scott | Exodus: Gods and Kings,The Counselor |

Figure 3.5.3: Average Titles – df.head(10)

| | Director | Flop |
|---|---|---|
| 136 | Tom Hooper | Cats,Cats |
| 325 | D.J. Caruso | The Disappointments Room |
| 467 | Guy Ritchie | Swept Away |
| 516 | Jan de Bont | Speed 2: Cruise Control |
| 694 | M. Night Shyamalan | The Last Airbender |
| 896 | Robert Rodriguez | The Adventures of Sharkboy and Lavagirl,Spy Ki... |
| 1164 | Dennis Dugan | Jack and Jill |
| 1299 | Michael Tiddes | Fifty Shades of Black |
| 1354 | Steven C. Miller | Escape Plan 2: Hades |
| 1360 | Sylvain White | Slender Man |

Figure 3.5.4: Flop Titles – df.head(10)

The figures 3.5.1 to 3.5.4, shows us the titles in the given dataset in terms of their box-office outcome. This is mapped by the correlation between HiddenGem Score and IMDb score, as these two features gives us the better idea whether the title was a box-office success or not.

### *F. Funnel Plot Representation – Country Wise Titles*

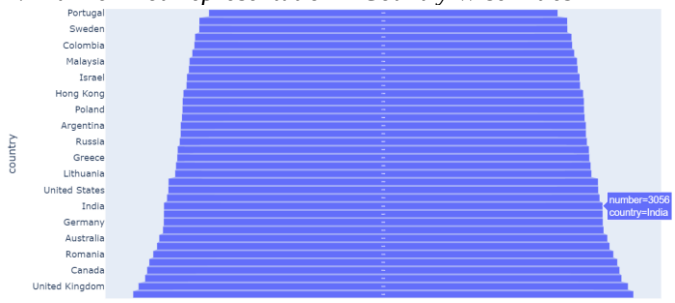



Figure 3.6.1: Country Wise titles

The figures 3.6.1 shows us the Funnel Plot representation of Country Wise titles. As we hover, we can see the number and country being displayed on the right side.

### *G. Geospatial Plot using Folium*

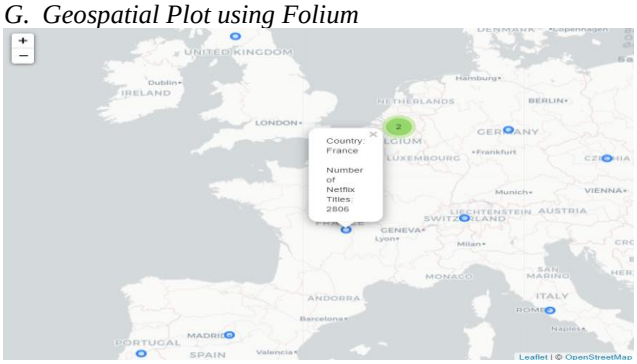



Figure 3.7.1: Folium Geospatial World Map

The figure 3.7.1, shows us the Folium Plot of World Map, where the above Funnel Plot data is mapped on a real map. As we hover and click on each country (those present in our dataset – 35 nos.) we get the Country name and Number of Netflix Titles. Using this we can get a Geospatial interface to visibly understand the country wise no. of titles.

## IV. SENTIMENT ANALYSIS

After performing EDA, we have decided to utilize and expand the outcome of the project by inculcating Sentiment Analysis based on the Series/Movie Reviews Data Set for

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

Training Set and Summary Attribute of Netflix Data as the Input/ Testing Data. As the above EDA catered to our pre-processing needs, where in we extracted the useful features, we have now combined three datasets together to form the training set, and we will use one of those three datasets as a testing set to obtain the classified data (result of our sentiment analysis).

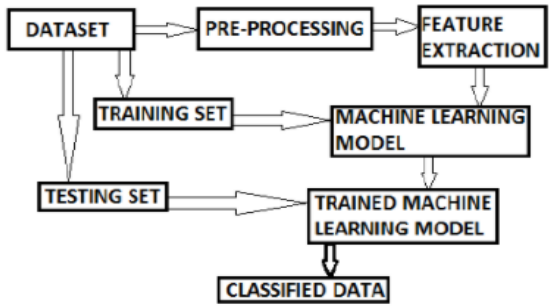



Figure 4.1: Sentiment Analysis Methodology.

### A. *Sentiment Counting*

```
positive    5028
negative    4973
Name: sentiment, dtype: int64
```

Figure 4.2: Count of Sentiments

The above figure 4.2 displays the no. of positive and negative sentiments present in the dataset.

### B. *Factorizing both the Sentiments*

```
(array([0, 0, 0, ..., 1, 0, 1], dtype=int64),
 Index(['positive', 'negative'], dtype='object'))
```

Figure 4.3: Sentiment Factorizing

Figure 4.3 shows the factorizing of the two sentiments as 0 and 1 i.e. for negative and positive.

### C. *Tokenizing the words present in Reviews*

```
{'the': 1, 'and': 2, 'a': 3, 'of': 4, 'to': 5, 'is': 6,
s': 13, 'as': 14, 'movie': 15, 'with': 16, 'for': 17,
'his': 24, 'have': 25, 'be': 26, 'one': 27, 'he': 28,
o': 35, 'from': 36, 'like': 37, 'or': 38, 'just': 39,
'there': 46, 'what': 47, 'some': 48, 'good': 49, 'when
```

Figure 4.4: Tokenizing words in reviews

Figure 4.4 tokenizes all the words in the Reviews Data set by uniquely identifying them as key and assigning a token number to each word.

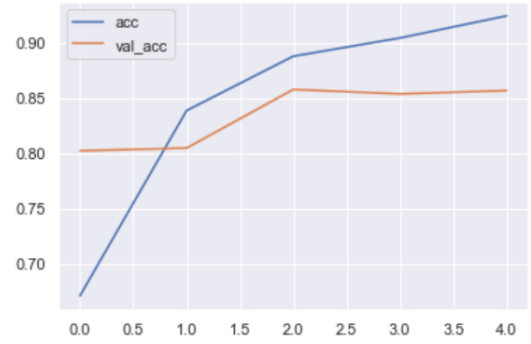
### D. *Training the model by fitting*

```
Epoch 1/5
250/250 [==============================]
y: 0.8026
Epoch 2/5
250/250 [==============================]
y: 0.8051
Epoch 3/5
250/250 [==============================]
y: 0.8581
Epoch 4/5
250/250 [==============================]
y: 0.8541
Epoch 5/5
250/250 [==============================]
y: 0.8571
```

Figure 4.5: Model Fitting

Once the tokenizing phase is completed the data generated is passed into the model for fitting. Thus the sentiment model is generated by training it using Keras.

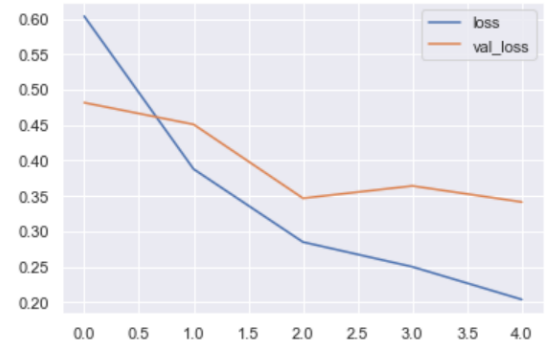
### E. *Plotting the Accuracy of the fit*




```
<Figure size 432x288 with 0 Axes>
```

Figure 4.6: Accuracy plot of the model

### F. *Plotting the Loss of the fit*




```
<Figure size 432x288 with 0 Axes>
```

Figure 4.7: Loss plot of the model

### G. *Output of the Sentiment Analysis*

```
Enter the movie name or its index to obtain Sentiment Analysis of its Summary:
Ong Bak 2: The Beginning

Movie/Series Sentiment Output:  positive elements present on a higher side.

Genre:  Action
```

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

## V. RESULTS

By utilizing the EDA approach towards the Netflix data we have garnered some crucial insights for other useful purposes. The following are summarized points:

- We have successfully cleaned the redudant records by removing them, and filtered the Data Set by discarding unused features.
- We developed a correlation amongst the utility features and established a guidance for our analysis.
- Built a plotting for size of Series & Movies in the dataset and also plotted IMDb_Scores based on their relevance count.
- Created a WordCloud – both unmasked and masked – based on the relevance of the Genres in the dataset.
- Insights based on SuperHit, Hit, Average and Flop box-office status of a Title using IMDb and Hiddden Gem Score as interlinked criteria, decided based on their correlation.
- Plotted the countrywise count of Netflix Titles using Funnel Plot and developed a Geospatial Plot using Folium based on the latter feature.
- Built a Sentiment Analysis Model by fitting the Series/Movie Reviews dataset, which obtains the result by making use of the summary column in the Netflix dataset.
- Displayed the accuracy and loss of the above model fitting as a plot.

Using the above methods and techniques we derived maximum results that are suitable for making better business decisions.

## VI. CONCLUSION

Data Analysis is a fundamental step to address the various needs of a client in any professional spectrum. The varied range of insights that can be derived from a data is itself primarily valuable in nature as there are multiple businesses that are actively looking for futuristic, predictive and descriptive insights from the already present raw data generated by them. It helps the organizations to gain access to numerous concealed patterns, information and bits of knowledge after the analysis had been performed. The analysis that we have just performed using the Netflix data not only provides us with incentives to take smart and intelligent business decisions, but also contribute to the overall growth of the firm. These insights maintain a clear sight and perspective for various stakeholders and help in targeting a positive vision for the future. The future scope of Data Analysis is bound to remain intact as long as businesses require Data Science in their everyday applicable decision-making processes. Also, there is a great scale of possibilities when it comes to developing unique interactive solutions and methods that are confined to make data exploration much more intriguing in nature. These constant advancements have stabilized a promising direction for data analysis as a systemic study that is going to stay as long as there is the crunch for data in any viable field of study in the real-world.

## VII. REFERENCES

[1] Kiranbala Nongthombam , Deepika Sharma, 2021, Data Analysis using Python, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 10, Issue 07 (July 2021)
[2] Jyoti Budhwar, Sukhdip Singh, 2021, Sentiment Analysis based Method for Amazon Product Reviews, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) ICACT – 2021 (Volume 09 – Issue 08)
[3] Soniya Grace, 2020, A Geospatial Analysis of Ground Water Quality Mapping using GIS in Sangareddy District, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 07 (July 2020)
[4] Gupta, Bhumika & Negi, Monika & Vishwakarma, Kanika & Rawat, Goldi & Badhani, Priyanka. (2017). Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python. International Journal of Computer Applications. 165. 29-34. 10.5120/ijca2017914022.
[5] https://en.wikipedia.org/wiki/Data_analysis
[6] https://towardsdatascience.com/represent-your-geospatial-data-using-folium-c2a0d8c35c5c