JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

An Approach of Data Visualization and Sentiment Analysis of Netflix classification using Machine Learning

Sukkala Rupika Sri[1], Dr Vanitha Kakollu[2]
PG Student[1], Assistant Professor[2]
Department of Computer Science, GITAM (Deemed to be University)
rsukkala@gitam.in,vkakollu@gitam.edu

ABSTRACT

In a recent situation, thanks to the pandemic OTT (Over The Top) becomes one of the most important entertainment online streaming. Netflix collects an enormous amount of knowledge because it's a really large subscriber base. This paper analyzes tons of knowledge and models from Netflix because this platform has consistently focused on changing business needs by shifting its business model from on-demand DVD movie rental and now focusing tons on the assembly of their original shows. a number of the foremost important tasks that analyzed from Netflix data are understanding what content is out there on Netflix. Does Netflix have more specialized TV Shows than movies in recent years? Recognize the similarities between the content and therefore the network between actors and directors and Understand what content is out there in several countries. to know sentiment analysis of content available on Netflix and the way the subscribers grow day by day. during this paper, I even have aimed to research the info using different parameters and present some data visualization using some python libraries like NumPy, Pandas, matplotlib, seaborn, and Machine learning.

KEYWORDS: Machine learning, Python, Data Visualization, Sentiment Analysis.

I. INTRODUCTION

Data visualization is essentially a graphical representation of knowledge and knowledge. The demand for data scientists is on the increase. Day by day we are shifting towards a memory world. it's highly beneficial to be ready to make decisions from data and use the skill of visualization to inform stories about what, when, where, and the way data might lead us to a fruitful outcome. Data visualization goes to vary the way our analysts work with data. it's getting to be expected to reply to issues sooner. And they'll get to be ready to dig for more insights check out data differently, more imaginatively. Data visualization is the advance creative data exploration. Our eyes are worn to colors and patterns. The user can quickly understand blue from yellow, circle from a square. Data visualization may be a sort of visual art that not only grabs our interests but also keeps our eyes on the message. It can narrate our entire numerical data to the stakeholders during a sort of captivating graphs with the assistance of knowledge visualization. immediately living in "an age of massive data" trillions of rows of knowledge are being generated a day. Data visualization helps us in curating data into a form that's easily understandable and also helps in highlighting a selected portion. Plain graphs are too boring for anyone to note and even fail to stay the reader engaged. Data Visualization is especially used for data cleaning, exploratory data analysis, and proper effective communication with business stakeholders. There are different stages that are:

In Data Cleaning means the method of deleting redundant columns, dropping duplicates, Cleaning individual columns, Removing the NaN values from the dataset of knowledge then modifying, replacing, or deleting them as required. Data Cleansing is taken into account as the essential element of knowledge

Science. Data Visualization is Using plots to find relations between the features. Data visualization is the graphical representation of data and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible thanks to see and understanding trends, outliers, and patterns in data. Word cloud is the process of graphical representations of word frequency that give greater prominence to words that appear more frequently in a source text. The larger the word within the visual the more common the word was within the document.

II. PROJECTED METHOD

Netflix is one of the foremost popular media and video streaming platforms. They have over 8000 movies or TV shows available on their platform, as of mid-2021, they need over 200M Subscribers globally. Netflix collects a huge amount of data because it has a very large subscriber base. In this way, it is very difficult to find the answer in records. so, we are using data visualization to display the data in the plot, bar, piechart, and Heatmap methods. Data visualization is an easy way to understand the data. At the last, understand sentiment analysis of content available on Netflix and it has millions of subscribers. The paper shows how many subscribers increasing day by day. In this paper, I have aimed to analyze the data using different parameters and present some data visualization using some python libraries like NumPy, Pandas, matplotlib, seaborn, and Machine learning.

III. SYSTEM ARCHITECTURE

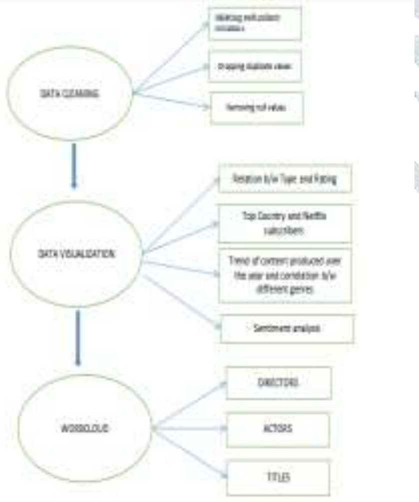

DATA CLEANING
Deleting redundant columns
Dropping duplicate values
Removing null values
DATA VISUALIZATION
Relation b/w Type and Rating
Top Country and Netflix subscribers
Trend of content produced over the year and correlation b/w different genres
Sentiment analysis
WORDCLOUD
DIRECTORS
ACTORS
TITLES

Figure-1: System Architecture

IV. METHODOLOGY

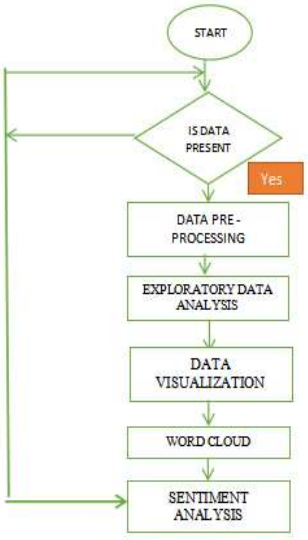


Figure-2: Methodology

Methodology: After data set extraction data can be analyzed and implemented in python and machine learning models. .Data pre-processing is performed to reduce the missing values and drop the duplicate values and remove the null values etc. Data Visualization is the process of analyzing the rating, type of movies and TV shows, Relation between the rating and type. Display the genres of Netflix and trends of content over the year. WordCloud is performed to display the graphical representations of word frequency. It analysis the top actor, director, and title in Netflix. then.Sentiment Analysis shows that the overall positive content is always greater than the neutral and negative content combined.

V. EVALUATION PROCESS

Netflix may be a popular entertainment service employed by people around the world. This Exploratory Data Analysis will survey the Netflix dataset through visualizations and graphs using python libraries, matplotlib, and seaborn. The dataset (netflix_titles.csv) and (Netflix subscribers.csv) consists of tv shows and movies available on Netflix as of 2022 and contains information including General information: id, title, type (TV Show or Movie), director, cast, and a brief description. Date fields: Firstly, data Deleting redundant columns after dropping duplicate values, then Cleaning individual columns and Removing the NaN values from the dataset. Secondly, data visualization shows the distribution of content rating on Netflix and type of movies and tv shows after analyzing the relation between type and rating. Recognizing what content is out there in several countries. Identifying similar content by matching text-based features. When the show was released and when it had been added to the catalog. This provides a unique perspective on the world's most lucrative cultural industry, reflected by Netflix: Movie and TV production.

VI. RESULT

Distribution of content rating on Netflix:

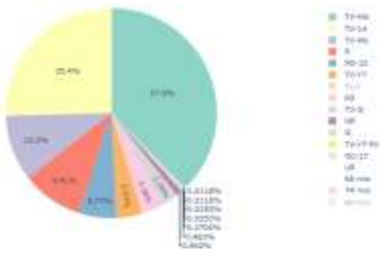


Figure-3: Rating on Netflix

The graph above shows that the majority of content on Netflix is categorized as “TV-MA”, which means that most of the content available on Netflix is intended for viewing by mature and adult audiences.

Type of Movie and TV show:

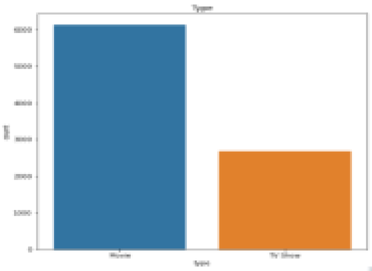


Figure-4: Type

Analysis entire Netflix dataset consisting of both movies and shows. Let’s compare the total number of movies and shows in this dataset to know which one is the majority.So there are about 6,000 movies and almost ,3000++ TV shows, with movies being the majority.

Relation between Rating and Type:

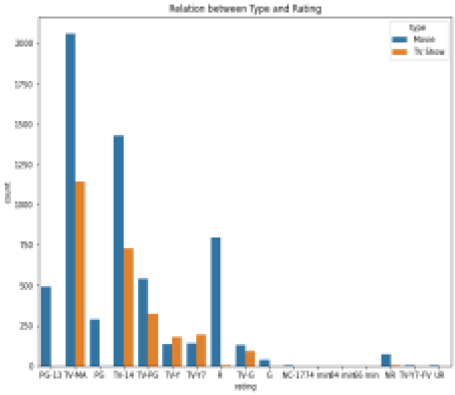


Figure-5: Relation between rating and type

The largest count of Netflix content is made with a “TV-14” rating. “TV-14” contains material that parents or adult guardians may find unsuitable for children under the age of 14. But the largest count of TV shows is made with a “TV-MA” rating. “TV-MA” is a rating assigned by the TV Parental Guidelines to a television program designed for mature audiences only.

Wordcloud using Python

Director:

Figure-6: Director

WordCloud shows the most popular director on Netflix, with the most titles, is mainly international.

Cast:

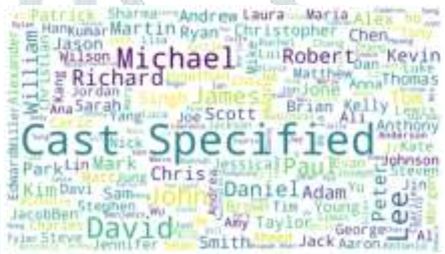


Figure-7: Cast

WordCloud shows the most popular actors on Netflix

Title:

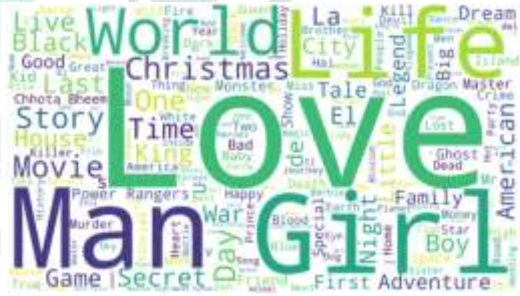


Figure-8: Title

WordCloud shows the most popular titles are used in the Netflix.

Top 10 Genres on Netflix:

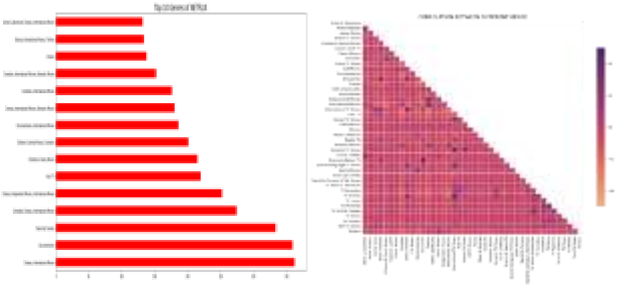
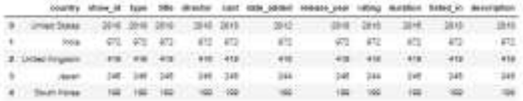


Figure-7: Genres on Netflix

From the graph, we know that International Movies take the first place, followed by action and adventures.

Which country has the most number of titles produced?

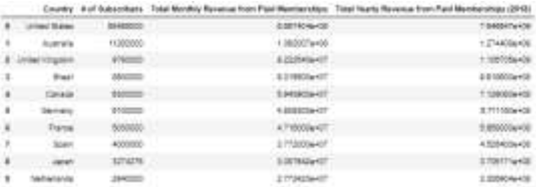


	country	show_id	type	title	director	cast	date_added	release_year	rating	duration	listed_in	description
0	United States	2818	2818	2818	2818	2818	2812	2818	2818	2818	2818	2818
1	India	972	972	972	972	972	972	972	972	972	972	972
2	United Kingdom	419	419	419	419	419	419	419	419	419	419	419
3	Japan	245	245	245	245	245	244	245	244	245	245	245
4	South Korea	199	199	199	199	199	199	199	199	199	199	199

Figure-8: Country

From the images above, we can see the top 15 countries contributor to Netflix. The country by the amount of the produces content is the United States.

Subscribers on Netflix:



	Country	# of Subscribers	Total Monthly Revenue from Paid Memberships	Total Yearly Revenue from Paid Memberships (2018)
0	United States	58460000	6.661404e+08	7.946847e+09
1	Australia	11300000	1.362007e+08	1.274408e+09
2	United Kingdom	9760000	9.222549e+07	1.105705e+09
3	Brazil	8800000	6.219800e+07	6.810000e+08
4	Canada	6500000	5.945800e+07	7.128000e+08
5	Germany	6100000	4.888300e+07	5.711100e+08
6	France	5050000	4.716000e+07	5.650000e+08
7	Spain	4000000	2.772000e+07	4.528400e+08
8	Japan	3274276	3.287842e+07	3.708171e+08
9	Netherlands	2840000	2.772423e+07	3.320904e+08

Figure-9: Netflix subscribers

Netflix has the large number of subscribers and it increasing day by day. 70M+ subscribers are in 2018 but Now, Netflix currently has 200+ million subscribers. Up from only 24.30 million subscribers in 2011. Netflix generated $24.99 billion in 2020. As of June 2021, Netflix has generated $14.5 billion in revenue thus far 2021. Netflix has 5,000+ content titles in their US library. 64.65% of Netflix subscribers are based outside of the United States.

Trend Of content over the years on Netflix:

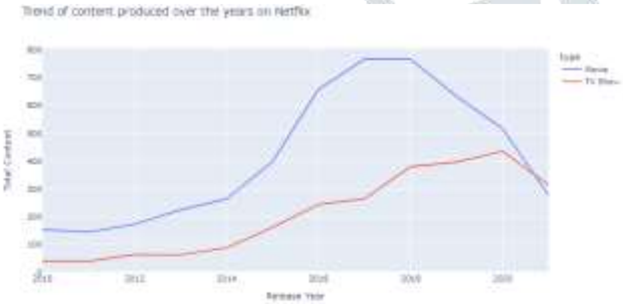


Figure-10: Content over the years

Based on the above, we can conclude that the popular streaming platform started gaining traction after 2013. Since then, the amount of content added has been increasing significantly. The growth in the number of movies on Netflix is much higher than that on TV shows. About 1000+ new movies were added in both 2019 and 2021. Besides, we can know that Netflix has increasingly focused on movies rather than TV shows in recent years.

Sentiment Analysis of Netflix:

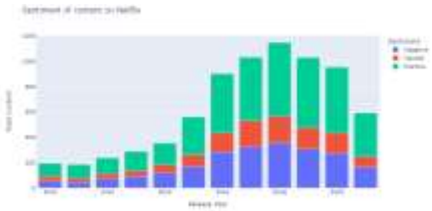


Figure-13: Sentiment Analysis

Sentiment Analysis shows that the overall positive content is always greater than the neutral and negative content combined.

VII. CONCLUSION

We have drawn many interesting inferences from the dataset Netflix titles; here's a summary of a few of them: The most content type on Netflix is movies and TV shows. The popular streaming platform started gaining friction after 2014. Since then, the quantity of content added has been increasing significantly, the country by the quantity of the produced content is that us, the foremost popular director on Netflix, with the most titles, is Rajiv chilaka. International Movies may be a genre that's mostly in Netflix, the most important count of Netflix content is formed with a "TV-14" rating, the foremost popular actor on Netflix TV Shows based on the number of titles is Takahiro Sakurai, The most popular actor on Netflix movie, based on the number of titles, is Anupam Kher, the most popular title on Netflix, based on the number of titles, is Love. Netflix mostly focused on dramas and international movies. Netflix has 200M+ subscribers in 2021 and finally, Sentiment Analysis shows that the overall positive content is always greater than the neutral and negative content combined.

VIII. REFERENCES

[1] K. Vanitha et al., The Development Process of the Semantic Web and Web Ontology.
(IJACSA) International Journal of Advanced Computer Science and Applications,, EBSCO Host, 2011, 2, 122-125

[2] GopaIyer, Aphrothiti J Hanrahan, Matthew I Milowsky, Hikmat Al-Ahmadie, Sasinya N Scott, Manickam Janakiraman, Mono Pirun, Chris Sander, Nicholas D Socci and Irina Ostrovnaya, "Genome sequencing identifies a basis for everolimus sensitivity," Science, Vol. 338, No. 6104, pp. 221– 229, 2012.

[3] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander and Joshua M Stuart, "The cancer genome atlas pan-cancer analysis project," Nature Genetics, Vol. 45, No. 10, pp. 1113– 1120, 2013.

[4] K. Vanitha et al., Opinion mining and Sentiment analysis of TELANGANA election on twitter data .Pramana Research Journal, Google scholar, 2014, 9, 147-154

[5] K. Vanitha et al., Implementing Multi prime RSA Algorithm to Enhance the Data Security in Federated Cloud Computing.
International Journal of Advanced Research in Computer and Communication Engineering, Google scholar, 2015, 4, 647-650

[6] Ashok, Meghana, et al, "A personalized recommender system using Machine Learning based Sentiment Analysis over social data," Electrical, Electronics and Computer Science (SCEECS), 2016 IEEE Students' Conference on. IEEE, 2016

[7] K. Vanitha et al., An Effective Stone Image Classification using Surface Patterns based on Reduced Dimension and Gray Level Range Model.International Journal of Advanced Research in Computer Science, Google scholar, 2017, 8, 1758-1765

[8] K. Vanitha et al., Sentiment Analysis of Election Result based on Twitter Data using R.International Research Journal of Engineering and Technology (IRJET), Google scholar, 2018, 5, 546-548

[9] K. Vanitha et al., Classification of Cancerous Profiles using Machine Learning Algorithms.
International Journal of Computer Trends and Technology (IJCTT), Google scholar, 2019, 67 , 99-101

[10] Soniya Grace et al., 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)

[11] Jyoti Budhwar et al., 2021, Sentiment Analysis based Method for Amazon Product Reviews, international journal of engineering research & technology (ijert) icact – 2021 (Volume 09 – Issue 08)

Sukkala Rupika Sri, pursuing Master of Data Science, Department of Computer Science, GIS, GITAM (Deemed to be University), Visakhapatnam. Her area of interest in Machine Learning.

Dr Vanitha Kakollu is currently working as Assistant Professor in the Department of Computer Science, GIS, GITAM (Deemed to be University). Her main areas of research include Image Processing, Data Mining and Machine Learning.