“Netflix - Data Visualisation and Recommendation System”

Ratna Pathak

Abstract— Netflix is the popular on-demand, over-the-top platform available today. With its presence in more than 190 countries. In this research project, a visual approach to explore and analyse the Netflix data set is applied and a graph-based recommendation system using the TF-IDF and Cosine similarity algorithm is done, using a combination of human reasoning and visual analytics approach. Despite the limitation of the recommendation system, it looks promising. The essential information about the content is visualised using python on observable notebook. Sentiment analysis is performed to get the overall polarity of the content over the years. Visualisations have revealed interesting data on the current underlying trend of the content delivered on Netflix.

1. Problem Statement

"Netflix is a subscription-based, video-on-demand, over-the-top streaming service and production company, Founded in 1997 by Reed Hastings and Marc Randolph in California, it offers a film and television series library through distribution deals as well as its productions, known as Netflix Originals.” The data set taken here is of Netflix with the information of contents available. This data was gathered with the help of a third-party application Flixable, a third-party search engine of Netflix, and made available on Kaggle. We will analyse the dataset to draw useful insights to investigate the research questions below, which will be helpful for the future release of content. With this, more users can be attracted to the platform, tending to spend most of their time watching shows and movies on Netflix.

The research questions are as follows:

1. Which age groups are targeted by Netflix in different countries? And which countries are the top content-producing countries?
2. Is there any best month for producers to add content and for subscribers to binge-watch the content?
3. Is there any trend of sentiments of the content over the years?
4. Can a recommendation system be made and visualised using the data available?

Answering the questions will be possible by doing an exploratory analysis of the dataset visually to understand the trends of movies and TV shows on Netflix, and finding answers to the questions will promote the development of similar content in future.

2. State of the Art

Netflix is the worlds leading Internet TV network with over 83 million members in over 190 countries enjoying more than 125 million hours of TV shows and movies per day, including original series, documentaries and feature films. Subscribers can watch on demand, on nearly any Internet-connected screen. Members can play, pause and resume watching, all without commercials or commitments. Netflix tries to give the best relevant recommendations to hook customers for a longer time.

In the existing literature, many works have tried to produce efficient recommendation engines using Term Frequency - Inverse Document Frequency and Cosine Similarity.

The three papers [1][2][3] I have analysed for this project, all vary in their approach to visual analytics and machine learning methods performed. Since visualisation is an important aspect of our project, I will differ by keeping things simple and maintaining a theme throughout, I will try to stick with the Netflix brand colour palette, which is mostly black, red and grey. The font family will be uniform across the visualisation, i.e 'Serif' and no interactive plots to avoid mess. The paper [2] also uses a geospatial plot using Folium to visualise the data on the map, I will try using a geodataframe for the same.

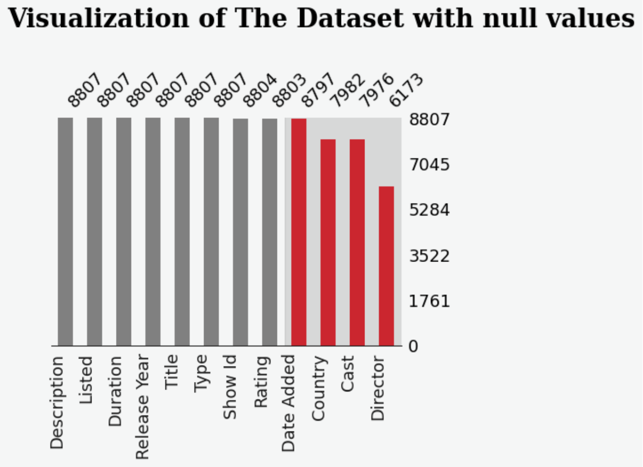
For the Recommendations system, (content-based) I will use the methodology of using TF-IDF and cosine similarity together to get the recommendation, I am going to base my report on this paper “Netflix Recommendation System based on TF-IDF and Cosine Similarity Algorithms”[1] and advance it further by doing k-means clustering on the vectorised description and obtain the related recommendation, also I will further visualise the relations on the graph using the networkx library of python. One of these papers “Exploratory and Sentiment Analysis of Netflix Data ”[2] used a detailed machine-learning approach for sentiment analysis, by adding additional data as sentiment, since sentiment analysis is not a major analysis here, I will try to use the basic database with doesn’t have sentiment feature but I will use a rule-based sentiment analyser, i.e, Textblob on the description of the content. The reference papers have tried using the word cloud for the title and description, for following the theme of Netflix I will do the word cloud mapped on Netflix’s N’ logo.

Analysing this paper[3], which has done some simple exploratory data analysis, I got much more ideas and possibilities for better visualisations of the data for gaining better insights into the research questions posed here. (411)

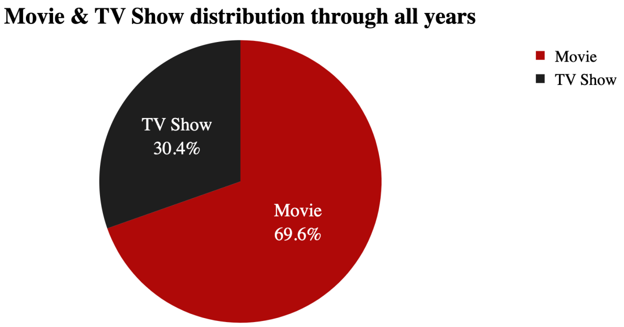
3. Properties of the Data

The data in hand is mostly categorical and time-dependent components. So it is better to convert some data frame columns to date-time format to extract month and year data from given data. Along with the categorical and numerical data, we do have plenty of text data from descriptions, so to analyse what kind of sentiment is prevailing over the past few years, Sentiment analysis is performed on the description's text data. To visualise the title's content, the Idea is to use the word cloud module to visualise that. In the below figure 1, we can see there are a total of 12 features, with 8807 entries where 4 features are having null values (highlighted in red), with 'director' having the most amount of null values and 'Date added' having a few.

Also to get an idea of the proportion of the type of content available is shown in the second figure in a pie chart.



The description of data and how the missing values were handled is given below:



- type - Two categories are there, most of the content on the platform is movies. Data is heavily skewed.
- director - Director names are there with some missing data, I'm not deleting or filling the missing data and keeping it as it is as it's not much of a problem in our analysis.
- cast- Actor/Actress names are there with some missing data and again I would keep the data as it is for the same reason.
- country - USA is by far the most content-producing country, so let's use this value to fill the null values.
- date_added - Jan 1 2020 is the most common date in content that went online on the platform. So, let's consider this to handle missing values.
- rating - This is a maturity rating, for telling which age group is appropriate to watch the content, so many categories are present, so it makes sense to find the main category, this is done with the help of 'Netflix help', and the main categories formed are as follows:

'TV-PG': 'Older Kids',
'TV-MA': 'Adults',
'TV-Y7-FV': 'Older Kids',
'TV-Y7': 'Older Kids',
'TV-14': 'Teens',
'R': 'Adults',
'TV-Y': 'Kids',
'NR': 'Adults',
'PG-13': 'Teens',
'TV-G': 'Kids',
'PG': 'Older Kids',
'G': 'Kids',
'UR': 'Adults',
'NC-17': 'Adults'

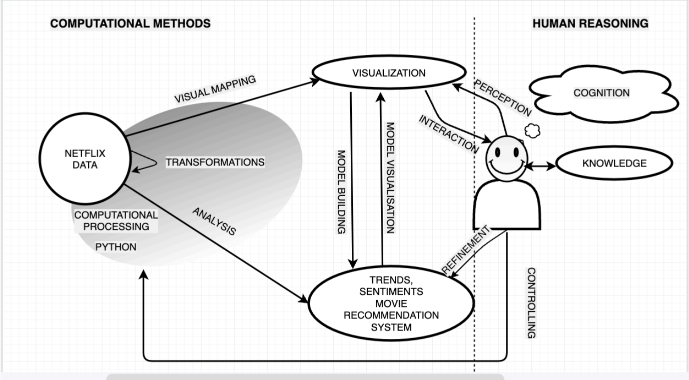
Now, this can be easily used to analyse the data to see how the audience is and what age group Netflix is targeting.

- listed_in - there are nearly 500 unique entries in this column, but we can split the data into some kind of string to see if we find anything interesting here.
- release_year - this gives the chronological timeline of produced content, here range of content is from 1925 to 2021.
- descriptions - this could help us understand what kind of topics the content is based on. Also can be used in finding the sentiments, and is Majorly useful in doing K-means clustering and applying TF-IDF and cosine similarity to get our recommendation system.

4. Analysis

4.1. Approach

The basic workflow is shown in figure 3 below, First is the data preparation stage: the type of data and missing values are analysed and filled accordingly. Then in the next stage: Visualisation is done using various libraries like Seaborn, Matplotlib, Plotly etc.



Next, Human cognition is applied to the visualisations to understand the underlying trends or patterns in data and derive insights from the data by applying human knowledge and reasoning.

With the insights and feedback from humans the data is processed into the required model, here is the recommendation system and Sentiment analysis and with this, the answers to research questions will also be tried to be found. All these processes are done using Python and its various libraries in observable.

Task 1.

Human: After dealing with the null value, the most important step in finding the answers to the research question is to convert the maturity rating available in the data into the basic understandable form of age group

Computational: In python, the basic age group is applied to reduce the category into a more useful one.

Human: For good visualisation selecting appropriate features as well as Visualization design is necessary. After many considerations, we will use a Heat map here. Which will give us the ability to identify targeted groups from various countries.

Task 2.

Computational: For knowing which is the best month for Netflix, content addition-wise, first the 'date added' needs to be separated into months and years.

Human: Deciding on the type of graph for Visualization, here we will try using a stacked pie chart, which will be used to observe the best month to add content and binge-watch.

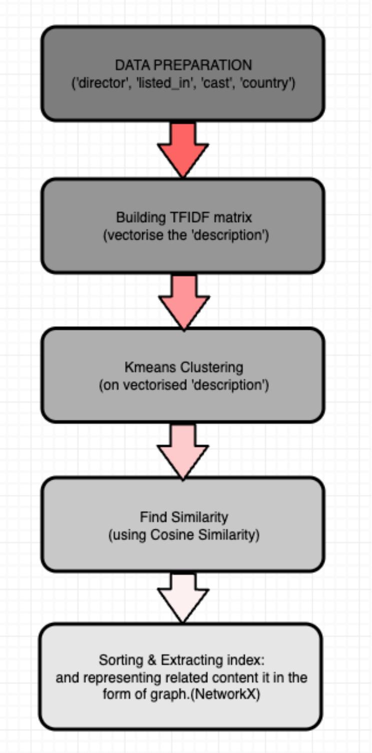
Task 3.

Computational: For sentiment analysis of the content description, Python's library TextBlob can be used.

Human: The use of a percentage bar chart can represent 3 polarities for the recent year and if any trend is there that can be observed by the human and a logical explanation can be provided by reasoning and doing research.

Task 4.

Computational: for making a recommendation system, heavy computation is required, in which steps involved are as shown in fig 3, I. Data Preparation: strip() function is applied on to remove unwanted things.


DATA PREPARATION
('director', 'listed_in', 'cast', 'country')
Building TFIDF matrix
(vectorise the 'description')
Kmeans Clustering
(on vectorised 'description')
Find Similarity
(using Cosine Similarity)
Sorting & Extracting index:
and representing related content it in the form of graph.(NetworkX)

II. TFIDF is applied to the description to vectorise it, to evaluate the importance of words.

III. K-means clustering is performed on the description to make 200 clusters.

IV. To find similarity Cosine similarity algorithm is used. Which measures the similarity by using the cosine angle between two vectors A and B, which can represent words.

V. Then sorting and extracting of the index is done to perform recommendations.

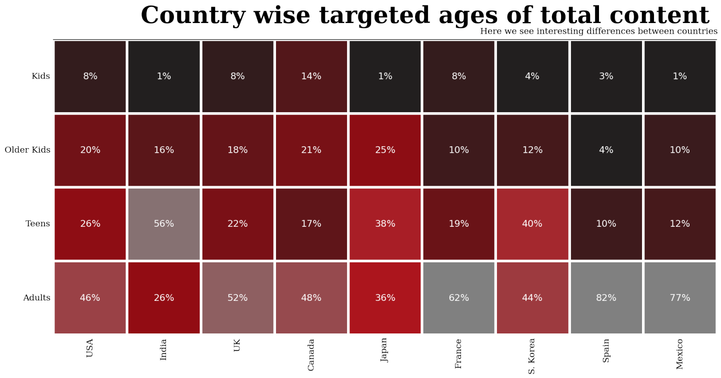
Fig. 3. Steps involved in making recommendations system.

To represent the relation graphically networkx is used. *Human:* The human analysis parts come to visualise the graphs with appropriate colours and sizes of several nodes.

4.2. Process

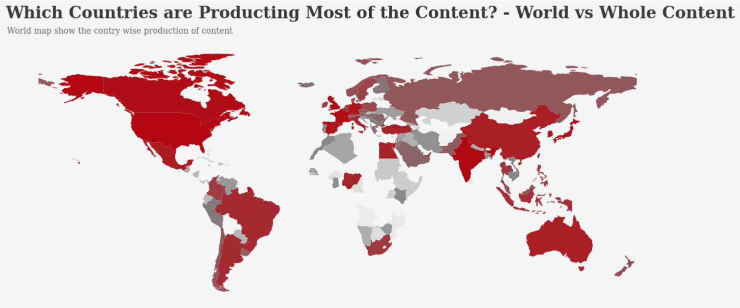
Everything in this project has been done in python using various libraries. My approach is to first understand Netflix's current trend in the type of offered programs by conducting exploratory data analysis and visualisations. Which will be the preparatory step for the second step of model building.

After doing EDA we know there is a vast difference in content preferences and active audience in different countries, The overall visualisation will not be useful, to be precise and see country-wise trends let's use heat map visualisation, let's focus on rating distribution entirely this time, by plotting age-order on y-axis and countries in the x-axis, and the intensity of colour will denote the %count of the age-order. We get the above



heat map:

Here we see interesting differences between the countries in targeting age group, As shown, India is leading in having the highest number of content for 'Teens' i.e, 56% of the total content, whereas Spain has the highest number of content made for 'adults' i.e 82% of the total content for adults. The typical age of Netflix users is younger adults and teenage bracket, a member of Gen Z or Millennials. Although one thing which might be a hindrance in getting exact data is Password sharing, with approx. 70% of Netflix consumers share their Netflix password. Netflix reports it loses $1.5 billion in potential income from password sharing. This is a very useful map for Netflix as well as content makers to analyse the target audience country-wise. It answers the first part of the research question. Now coming to the second part :

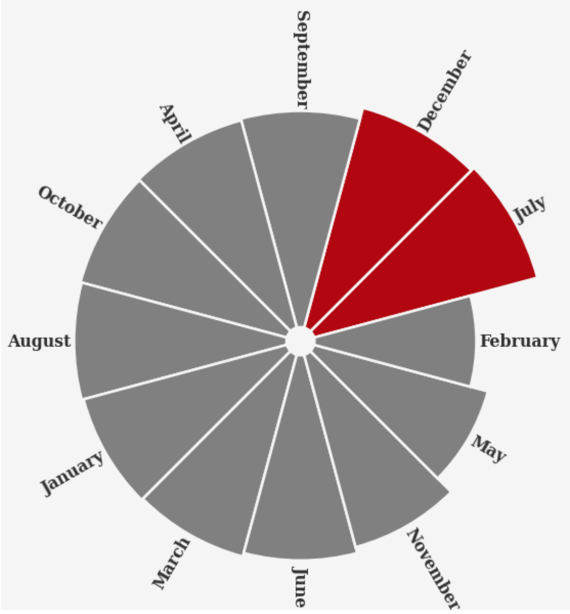

Which Countries are Producting Most of the Content? - World vs Whole Content
World map show the contry wise production of content

The country data is visualised on the geodataframe of python and visualised Dark red colour of the country indicates the highest content-producing country and as the colour fades to white, is the least content-producing country in the world.

It gives us a clear idea that most of the contents are produced by the US followed by India, as seen in bright red. Netflix subscriber count in the Asia-Pacific region has grown the most in the last two years, but this remains its least penetrated region, even setting aside its inability to offer the service in China. Also, there is a difference in the pricing and offers. In comparing the regional numbers with the populations of the regions, significant opportunity for growth exists in all but the US and Canada regions. Netflix is reasonably described as having multinational availability, and significant take-up around the globe. Content Produced in particular locations helps Netflix to make content strategies and priorities. Such data is very useful for business analysis

For the 2nd research question, the best month to add content can be obtained by computing the maximum and a minimum of added_month (obtained from the date added), computing the width of each bar and computing the angle at which each bar is centred.

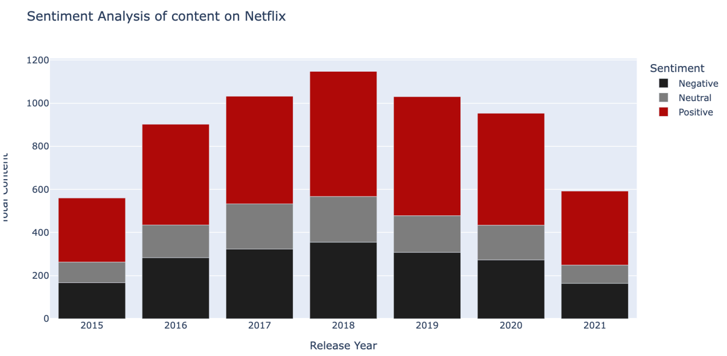
From the highlighted part of the pi chart shown in the fig, it is clear that the month of December and July is the month in which most of the content went online. It is good to observe that most of the content starts to be available in the holiday season, like in December. Which is Christmas and New year's eve. Although for July I can not see any major holiday but may be this month is randomly kept in between the year to keep subscribers entertained in between.



So, December and July are the best months for producers to add and subscribers to binge-watch new content on the platform.

For 3rd Research question: For the sentiment analysis of the description, Python's library TextBlob can be used. Basically, TextBlob is a library for processing textual data. TextBlob is a Lexicon-based sentiment analyser, It has some predefined words and a weighted dictionary, where it has some scores to help to calculate a sentence's polarity.

From the below bar graph of sentiment analysis of recent years, we can see that it's in a pattern following a curve, the sentiments in general increase from 2015 to 2018 and then decreases. The overall positive content is always greater than negative and neutral content. The decline from the year 2019 may be because of the covid affecting the industry. The dominance of positive data shows may be the viewer likes to watch positive content much more than neutral or negative content.

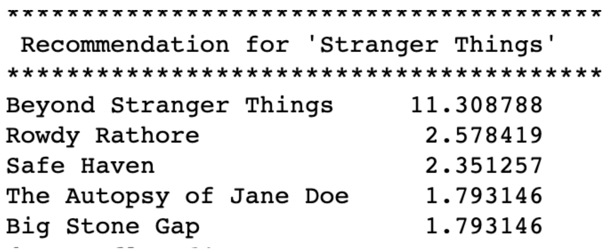


For the last, (4th) research question of making a recommendation system:

In general recommendation systems are of three types.

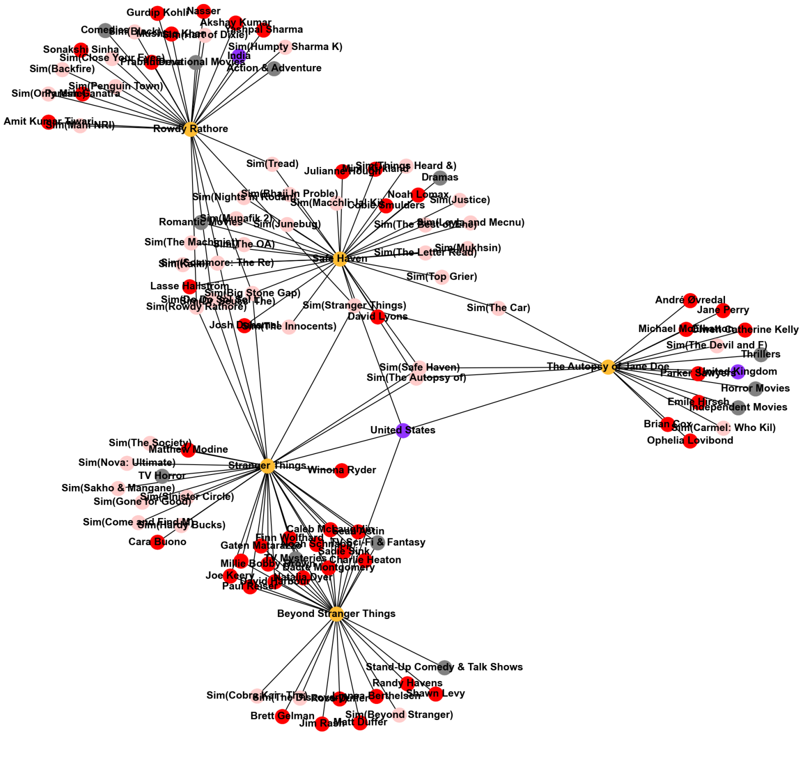
Trend-based recommendation systems - These are like counting views per video, clicks, and upvotes in the last few hours. *Content-based recommendation systems* - These are purely content-dependent recommendations, like songs with a similar genre, singer, director, country and so on are used in finding the similarities. Similarities are analysed and found within the products or articles. *User-User recommendation systems* - These are adaptive systems, which collects huge users' data and find a correlation between users' search trends, activities, and watch lists and suggests similar content for similar kind of users. This is the one which big companies like Netflix use.

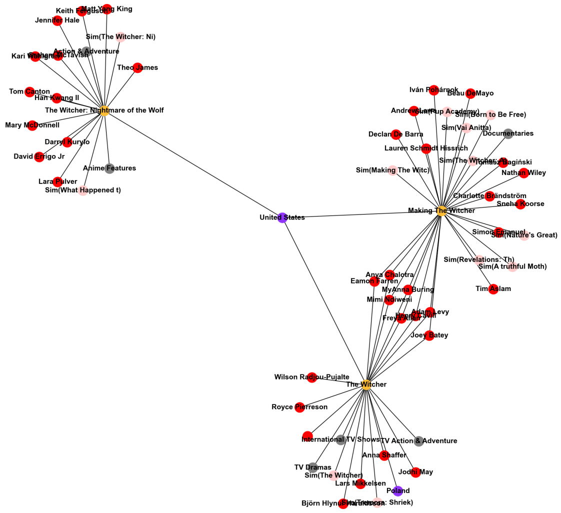
Here I am attempting to implement a simple content-based recommendation system with the available data. Idea is to concatenate the whole content data such as title, director, age rating, description, and country together to create text for each movie or show. And then, vectorise this using TF_IDF, Then the vectorised data is clustered(k-means) into 200 different categories and pass it through a Cosine Similarity method to get the angular distances between various content and create a similarity matrix. The similarity matrix will give the extent of similarity between each content and sorting and extracting the index can give the recommendations. For the series of 'stranger things' these recommendations were given: Beyond Stranger Things, Rowdy Rathod, Safe Heaven, The Autopsy of Jane Doe and Big Stone Gap.



```
****************************************
 Recommendation for 'Stranger Things'
****************************************
Beyond Stranger Things      11.308788
Rowdy Rathore                2.578419
Safe Haven                   2.351257
The Autopsy of Jane Doe      1.793146
Big Stone Gap                1.793146
```

The number beside the recommended movie suggests the intensity of the relationship that the movie has with the stranger things. We can visualise one more group of content to understand the graph better: yellow dots denote the content and purple denotes the country. This graph shows the relatability of one content with another. Below is an example of a relationship between the movie: The Witcher, Making the Witcher and The Witcher: Nightmare of the wolf, we can see two of them share the casts but the nightmare of the wolf doesn't have much in common.





4.3.Results

- We were able to find that different age groups are targeted by Netflix in different countries.
- Also with the help of a map we were able to distinguish the most content-making countries i.e, the US followed by India, observing different content produced in particular locations helps Netflix to make content strategies and customer prioritisation.
- We successfully found two months when most of the content was added. July and December were found to be the best months for producers to add content and for subscribers to binge-watch new content added. For December the most probable reason for being a holiday month was considered, but July, being one of the top months, surprisingly, could not

find much logical explanation.

- We analysed the trend in sentiment analysis and found that most of the content present is positive indicating the preference of the subscribers. Also, there was a curve pattern observed but that was most probably because of covid affecting the production industry.
- And finally, we were able to implement a working content-based recommendation system based on TF-IDF and cosine similarity, by applying k-means clustering, and visualising it using the networkx.

5. Critical reflection:

- Data cleaning and processing help us to beautifully visualise and draw insights making it the most important step.
- The final visualisation was achieved after visualising different features in various different graph formats and then selecting a more insightful one.
- We know how analytics is used for recommendation systems to improve the customer experience. Hence it's one of the major topics in the subscription-based platform. A working content-based recommendation system was implemented here. Which can be improved. The relevance of the results of this system can be criticised because of the limited number of occurrences present in the used texts. Although, it gives the advantage of being easily implemented and used. Big companies like Netflix, Amazon Videos, and Spotify use great recommendation systems which are adaptive and very efficient as they are powered by a large amount of subscribers' data available.
- The graphical representation of the recommendation system gives a great insight into how the movies are related to each other and with what intensity.
- Analysis of the month in which most data is added, We successfully found two months when most of the content was added. July and December were found to be the best months. For December the most probable reason for being a holiday month was considered, but July, being one of the top months, surprisingly, could not find much logical explanation. Also for producers to add new content to these months might not be as useful. It can not be just derived from this factor as it might happen that with so much content added, it might not get many views due to the presence of a large number of content. So a bit more analysis is required to get to the conclusion for the producer part but for subscribers to binge-watch it is always great to take a Netflix subscription when there is a lot of new content added to binge-watch.
- Sentiment Analysis of the content shows that the overall positive content is always greater than the neutral and negative content combined. But this also could be dependent on what is trending, and it would be interesting to know more details of sentiments of the content by country, genre, casts etc.
- Analysing Netflix data not only provide incentives to take smart and intelligent business decisions but also contributes to the overall growth of the firm. These insights maintain a perspective for various stakeholders and help in targeting a positive vision for the future.
- Here is a word cloud visualisation of the Netflix logo, with the content of its title showing frequently used words are Love, Christmas, girl, etc. this can give us insights into what titles are getting more attention.

Love
Christmas
Girl
Man
Day
World
Life
Movie
Story
Last
Time
Black
Secret
One
King
American
House
Boy
Game
City
Live
Little
War
Dream
Night
Family
La
El
de
First
Good
Tale
Adventure

Table of word counts

Problem statement	232
State of the art	411
Properties of the data	480
Analysis: Approach	462
Analysis: Process	1058
Analysis: Results	187
Critical reflection	443

References

1. Mohamed Chiny, Marouane Chihab, Omar Bencharef and Younes Chihab ‘Netflix Recommendation System based on TF-IDF and Cosine Similarity Algorithms’ BML 2021
2. Sukkala Rupika Sri[1], Dr Vanitha Kakollu. ‘An Approach of Data Visualization and Sentiment Analysis of Netflix classification using Machine Learning,2022 JETIR’.
3. KarthikBabu Vadloori, Shriya Madhavi Sanghishetty. ‘Exploratory and Sentiment Analysis of Netflix Data. ‘International Journal of Engineering Research & Technology (IJERT), 2021.
4. Karthik Srivatsa Maddodiand Krishna Prasad. K, ‘Netflix Bigdata Analytics- The Emergence of Data-Driven Recommendation’, (IJCSBE), ISSN: 2581-6942, Vol. 3, No. 2, October 2019.
5. Kiranbala Nongthombam, Deepika Sharma, 2021, Data Analysis using Python, INTERNATIONAL JOURNAL OF

ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 10, Issue 07 (July 2021)

6.Bennett, J., Lanning, S., 2007. The Netflix prize. In Proceedings of KDD cup and workshop, Vol. 2007, p. 35). New York, NY, USA. Chiny, M., Bencharef, O., Hadi, M.-Y., Chihab, Y., 2021. A Client-Centric Evaluation System to Evaluate Guest's Satisfaction on Airbnb Using Machine Learning and NLP. Ap

7.Chiny, M., Bencharef, O., Hadi, M.-Y., Chihab, Y., 2021. A Client-Centric Evaluation System to Evaluate Guest's Satisfaction on Airbnb Using Machine Learning and NLP. Applied Computational Intelligence and Soft Computing.

8. Source of dataset :https://www.kaggle.com/datasets/shivamb/netflix-shows