International Journal of Research in Science & Engineering
ISSN: 2394-8299
Vol: 04, No. 02, Feb-Mar 2024
http://journal.hmjournals.com/index.php/IJRISE
DOI: https://doi.org/10.55529/ijrise.42.19.26

Content Based Recommendation System on Netflix Data

Dr. Deepti Sharma[1], Dr. Deepshikha Aggarwal[2*], Dr. Archana B. Saxena[3]

[1,2,3]Professor, Department of Information Technology, JIMS, Sec-5, Rohini, Delhi, India.*

Corresponding Email: [2]deepshikha.aggarwal@jimsindia.org*

Received: 28 September 2023 **Accepted:** 17 December 2023 **Published:** 01 February 2024

Abstract: ***After pandemic, OTT platforms are the most common platform to provide entertainment to users. Among all platforms, Netflix has become most the popular one. Data visualization of Netflix data can provide valuable insights and benefits in many ways like understanding viewer preferences, content optimization, personalized recommendation, quality and content performance evaluation, fraud detection to name a few. This research provides exploratory data visualization and provide a content based recommendation system on Netflix data as in real world applications, company uses these recommendation system algorithms to determine which system are better to improve users' engagement of the platform.***

Keywords: Data Set, Data Visualization, Content Based System.

1. INTRODUCTION

After pandemic, OTT platforms are the most common platform to provide entertainment to users. OTT stands for "Over-the-Top." It refers to streaming services that deliver content over the internet, bypassing traditional cable or satellite television providers. These services allow users to stream movies, TV shows, and other video content directly to their devices, such as smartphones, tablets, smart TVs, and computers, via the internet. Examples of popular OTT platforms include Netflix, Amazon Prime Video, Hulu, Disney+, and HBO Max. OTT services have become increasingly popular due to their convenience and the ability to access a wide range of content on-demand, anytime and anywhere with an internet connection.

Among all platforms, Netflix has become most popular one because of many reasons. Firstly, it has a vast content library of movies, TV shows, documentaries, and original content. Subscribers have access to a wide range of genres and languages, catering to various tastes and preferences. Secondly, Netflix invests heavily in producing original content, including movies,

International Journal of Research in Science & Engineering
ISSN: 2394-8299
Vol: 04, No. 02, Feb-Mar 2024
http://journal.hmjournals.com/index.php/IJRISE
DOI: https://doi.org/10.55529/ijrise.42.19.26

TV series, and documentaries. Some of these original productions have gained critical acclaim and have a dedicated fan base. Examples include "Stranger Things," "The Crown," and "The Witcher". Another advantage is that Netflix provides a user-friendly interface that is easy to navigate. It offers personalized recommendations based on viewers' watch history and preferences, making it easier for users to discover new content they might enjoy. Netflix is compatible with a wide range of devices, including smartphones, tablets, smart TVs, laptops, and gaming consoles. Netflix allows users to download selected movies and TV shows for offline viewing. This feature is especially popular among commuters and travellers, as it enables them to watch content without an internet connection. Netflix is available in numerous countries around the world, offering content in multiple languages. Its global presence allows it to cater to a diverse audience with a variety of cultural and regional content. These factors, combined with effective marketing strategies and a focus on customer satisfaction, have contributed to Netflix's popularity as a leading streaming service.

2. RELATED WORK

Data visualization of Netflix data can provide valuable insights and benefits in various ways. By visualizing data on what content is being watched, when, and how frequently, Netflix can gain insights into viewer preferences. This information helps in making data-driven decisions regarding content creation, acquisition, and recommendations. Data visualization allows Netflix to analyze which genres, actors, directors, or specific themes are popular among viewers. This information can be used to optimize their content library, creating more shows and movies that align with audience interests. Netflix uses algorithms to recommend content to users based on their viewing history. Data visualization helps in understanding how effective these recommendations are, allowing for continuous improvement. By visualizing user behavior patterns, Netflix can refine its recommendation algorithms for a more personalized user experience. Visualization of user engagement data, such as watch time, can help Netflix identify trends and patterns. For instance, they can analyze when users are most active, helping them plan releases or promotional activities during peak usage hours. Netflix can visualize data related to video quality, buffering rates, and user feedback. This information helps them ensure a seamless streaming experience for users by identifying and resolving issues promptly. By visualizing data on viewer ratings, reviews, and social media discussions, Netflix can assess the performance of their original content. This feedback can guide decisions about renewing shows, creating sequels, or investing in similar projects. Data visualization provides a holistic view of Netflix's business performance, including subscriber growth, churn rates, regional preferences, and revenue patterns. This information is crucial for strategic planning, marketing initiatives, and expansion efforts. Visualization techniques can be applied to detect unusual viewing patterns that might indicate fraudulent activities, such as account sharing or unauthorized access. This helps Netflix maintain the integrity of its subscription model.

In summary, data visualization of Netflix data is essential for making informed decisions, improving user experience, enhancing content offerings, and ensuring the overall success and sustainability of the streaming platform.

International Journal of Research in Science & Engineering
ISSN: 2394-8299
Vol: 04, No. 02, Feb-Mar 2024
http://journal.hmjournals.com/index.php/IJRISE
DOI: https://doi.org/10.55529/ijrise.42.19.26

3. METHODOLGY

One of the commonly used datasets related to Netflix is the "Netflix Movies and TV Shows" dataset available on Kaggle. This dataset provides information about movies and TV shows available on Netflix as of 2021. You can find it on Kaggle's website by searching for "Netflix Movies and TV Shows" dataset.
The Netflix films and TV series that are accessible as of 2019 are included in this dataset. The third-party Netflix search engine Flixable is where the dataset was gathered.
An intriguing analysis was issued where quantity of TV series available on Netflix has almost tripled. Since 2010, the number of films available on the streaming service has dropped by over 2,000, but the number of TV series has increased by almost three times. Investigating what further insights might be gleaned from the same dataset will be interesting.

Exploratory Data Analysis and Methodology
A. Understanding What Content is Available in Different Countries
Currently Netflix is available in 113 countries and regions. Let's see which country has the most content.

Fig 1: Different content available in different countries

Above figures observe that US, India and UK are at position first, second and third respectively out of top 10 countries with most content on Netflix.

International Journal of Research in Science & Engineering
ISSN: 2394-8299
Vol: 04, No. 02, Feb-Mar 2024
http://journal.hmjournals.com/index.php/IJRISE
DOI: https://doi.org/10.55529/ijrise.42.19.26

B. Determine if Netflix has Increasingly Focusing on TV Rather than Movies in Recent Years.

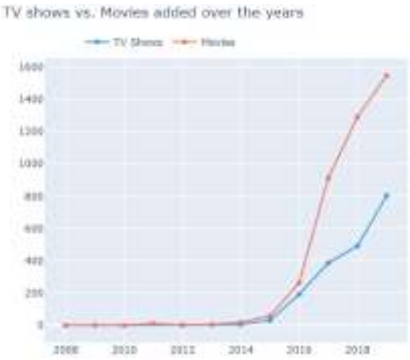

Fig 2: TV shows vs. Movies popularity

C. Year Wise Distribution and Type of Content Added to Netflix

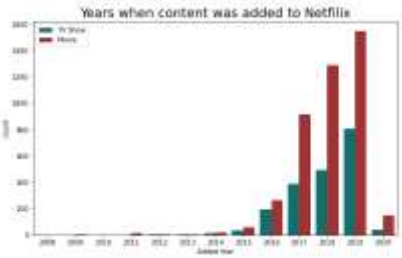

Fig 3: Content Addition

D. Network analysis of Actors / Directors

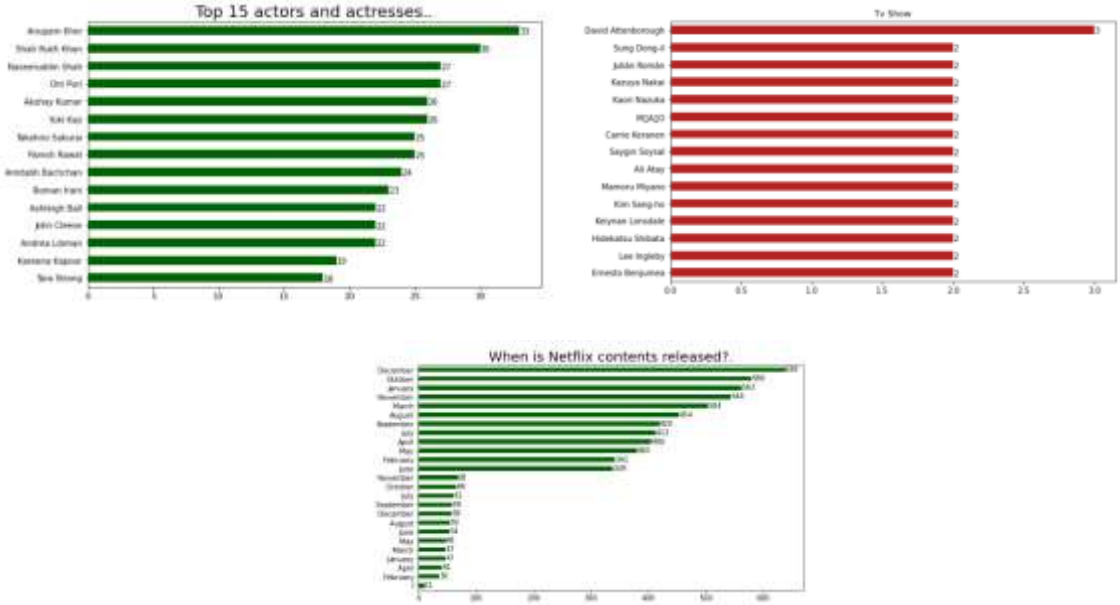

Fig 4: Network Analysis of Actors/Directors

International Journal of Research in Science & Engineering
ISSN: 2394-8299
Vol: 04, No. 02, Feb-Mar 2024
http://journal.hmjournals.com/index.php/IJRISE
DOI: https://doi.org/10.55529/ijrise.42.19.26

Evaluation Process

One of Netflix's main business operations is providing viewers with tailored show recommendations. To do this, they have created proprietary, complex recommendation system. This recommendation engine works hard to make it as easy as possible for you to choose a show or movie to enjoy whenever you use the Netflix service. They calculate the probability that you will view a specific title in the catalogue depending on several variables, such as:

- Your experiences using our service (including the films you've watched and the titles you've rated)
- Other users of our service who share similar interests and preferences.
- Details about the titles, including the actors, categories, release year, and genre.

They look at things like the following in addition to what you have viewed on Netflix in order to best tailor the recommendations: Depending on when you watch, the gadgets you're using to stream Netflix, and length of your viewing. They process each of these data sets as inputs into algorithms. Demographic data, including age or gender, is not taken into consideration while creating recommendations by the recommendations system.

Netflix system uses sophisticated algorithms and complicated processes to deliver a personalised experience. It not only selects which titles to display in the rows on your Netflix homepage, but it also ranks each title within the row. Finally, it ranks the rows themselves. Stated differently, their technologies have arranged the titles in your Netflix site so that they are presented in the best possible order for you to enjoy.

There are three levels of customisation in each row:

- The row selection (e.g., Keep Watching, Hot Right Now, Best Comedy, etc.)
- Titles that show up in the row, and
- The order of those titles

Every time you use the Netflix service, they gather input from you and use that information to continuously retrain their algorithms to forecast what you're most likely to watch with greater accuracy. The computation systems, algorithms, and data keep feeding back on each other to generate new suggestions so that you can get a product that makes you happy.

4. RESULTS AND DISCUSSIONS

Content Based Recommendation of Netflix Shows

One of Netflix's main business operations is providing viewers with tailored show recommendations. The Netflix recommendation system operates at a very high level and enumerates several elements that the streaming service employs to build its suggestion system.

A. Viewers-Level Factors

Customers' watching history, the ratings they have given other shows, when they use Netflix, how long each active session lasts, the device they use to watch shows on, and other Netflix users who have similar interests.

International Journal of Research in Science & Engineering
ISSN: 2394-8299
Vol: 04, No. 02, Feb-Mar 2024
http://journal.hmjournals.com/index.php/IJRISE
DOI: https://doi.org/10.55529/ijrise.42.19.26

B. Text-Based Features of the Shows

Release year, generation, content etc. Since there don't seem to be any views-level characteristics in the dataset, we will mostly use text-based features to construct our recommendation system. Here, the objective is to suggest five shows based on a list of shows that viewers have already seen using the following features:

- director
- description
- listed in
- rating

Data Preparation

The four variables—director, description, listed in, and rating—are concatenated to generate the variable aggregated_text, which is the first thing we do in this section. Next, we convert every word in our corpus to lowercase. Take out the English stopwords and punctuations.

```
df_netflix = df_netflix[df_netflix['title'].notna()]
df_netflix['aggregated_text'] = df_netflix['description'].str.lower() + " "
+ df_netflix['listed_in'].str.lower() + " " + df_netflix['rating'].str.lowe
r() + df_netflix['director'].str.lower()
corpus_tokenized = list(df_netflix['aggregated_text'].str.split(" "))
stopwords_list = set(stopwords.words("english"))
index = list(range(0, len(corpus_tokenized)))
clean_corpus = []

for sentence in corpus_tokenized:
  s = []
  for word in sentence:
    clean_word = re.sub(r'[^\w\s]','', word)
    if clean_word not in stopwords_list:
      s.append(clean_word)
  clean_corpus.append(" ".join(s))
```

Fig 5: Data Preparation

TF-IDF Vectorizer

Our corpus is vectorized using TF-IDF vectorizer, an acronym for Term Frequency-Inverse Document Frequency, which converts the unprocessed text into a matrix of TF-IDF features. I prefer TF-IDF vectorizer over CountVectorizer since word counts do not account for word frequency in different documents. Certain words (e.g., "man") may appear several times in a show's description; nonetheless, their big counts will have little value in the encoded vectors.
To determine how similar two episodes are to one another in terms of text-based elements like content, rating, genres, and directors, we employ a metric called cosine similarity. The vector representation of their text-based attributes is more comparable the higher the cosine similarity.

```
tfidf_vectorizer = TfidfVectorizer().fit_transform(clean_corpus)
```

International Journal of Research in Science & Engineering
ISSN: 2394-8299
Vol: 04, No. 02, Feb-Mar 2024
http://journal.hmjournals.com/index.php/IJRISE
DOI: https://doi.org/10.55529/ijrise.42.19.26

5. CONCLUSION

For evaluating the above algorithm, let's take an example. Suppose I have watched three TV shows on Netflix: Stranger Things, The Vampire Diaries, Sense8. Let's find out what are the 5 shows that Netflix will recommend to me?

```
watched_shows = ['Stranger Things', 'The Vampire Diaries', 'Sense8']
get_recommendation(watched_shows, tfidf_vectorizer)
```

	Title	Genre	Cosine_similarity
0	Anjaan: Special Crimes Unit	International TV Shows, TV Horror, TV Mysteries	0.222666
1	The OA	TV Dramas, TV Mysteries, TV Sci-Fi & Fantasy	0.221037
2	Haven	Classic & Cult TV, TV Horror, TV Mysteries	0.219799
10	Skins	British TV Shows, International TV Shows, TV D...	0.214939
3	Star-Crossed	Romantic TV Shows, TV Dramas, TV Sci-Fi & Fantasy	0.209957

Fig 6: Recommendation System

Thus in real world applications, company uses these recommendation system algorithms to determine which system are better to improve users' engagement of the platform.

6. REFERENCES

1. Sharma D., Saxena B. A., Aggarwal D. "Exploratory Sentiment Analysis of Sales Data", "European Economic Letters" (ABDC Journal, C Category), ISSN: 2323-5233, Vol 13 No. 4, October 2023, Page no: 982-986.
2. G. Adomaviciu and A. Tuzhilin, "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions," IEEE, vol. 17, no. 6, pp. 734-749, 2005
3. G. Linden, B. Smith and J. York, "Amazon.com recommendations: item-to-item collaborative filtering," IEEE Internet Computing, vol. 7, no. 1, pp. 76-80, 2003.
4. J. Lu, D. Wu and W. W. G. Z. Mingsong Mao, "Recommender system application developments: A survey," Decision Support Systems, vol. 74, pp. 12-32, 2015.
5. "Movie Dataset: Budgets, Genres, Insights," [Online]. Available: https://www.kaggle.com/datasets/utkarshx27/movies-dataset. [Accessed 3 March 2023].
6. S. Prakash, A. Nautiyal and M. Prasad, "Machine Learning Algorithms for Recommender System - a comparative analysis," International Journal of Computer Applications Technology and Research, vol. 6, no. 2, pp. 97-100, 2017.
7. X. Yang, K. Yang, T. Cui, M. Chen and L. He, "A Study of Text Vectorization Method Combining Topic Model and Transfer Learning," Processes, vol. 10, p. 350, 2022.
8. Sharma D., Saxena BA., "Exploratory Data Analysis and Visualization: Netflix data Using Python Libraries", International Journal of Creative Research Thoughts (IJCRT), ISSN:2320-2882, Volume.11, Issue 10, pp.f296-f301, October 2023

International Journal of Research in Science & Engineering
ISSN: 2394-8299
Vol: 04, No. 02, Feb-Mar 2024
http://journal.hmjournals.com/index.php/IJRISE
DOI: https://doi.org/10.55529/ijrise.42.19.26

9. Devesh Lowe, Bhavna Galhotra, Yukti Ahuja, "Discovering Binge watching and Audience Engagement through Sentiment Analysis", International Journal of Advanced Science and Technology, ISSN 2005 4238, Vol. 29, No. 7, (2020), pp. 8030-8038
10. [10] Ronak Sharma, Devesh Lowe, Bhavna Galhotra,"A Study on Lexicon Based Techniques of Twitter Sentiment Analysis", IEEE, August 2022. DOI: 10.1109/ICACCS 54159.2022.9785231