A Research Paper On

OTT Analytics

Submitted By

Name : Swaroop U. Shinge

Roll No: 22201232

MCA SY

Name : Vaibhav B. Gaikwad

Roll No: 22201234

MCA SY

Under Guidance Of

Dr. P.P. Shinde

Introduction

Over-the-top (OTT) relates to the transmission of audio, video, and various other media content via the internet instead of traditional cable or satellite distribution. OTT platforms like Netflix, Amazon Prime Video, and Hulu provide a variety of content to viewers immediately through subscription-based or ad-supported channels.

The term of over-the-top (OTT) entertainment obtained traction in the early 2000s, because of largely to the development of internet technologies and the wide availability of fast internet connections. This technology allowed the efficient online streaming of excellent in quality video and audio content. The growth of OTT was additionally impacted by changing viewer choices, as audiences looked for on-demand and individualized experiences.

In the mid-2000s, Netflix transitioned from a DVD rental service to a streaming platform, as well as significant advancements in streaming technology, such as video compression and content delivery networks (CDNs). Because of these advancements, a global audience can now enjoy smooth and uninterrupted streaming experiences.

The media and entertainment landscape has been transformed by OTT platforms. They've expanded globally, changed viewer habits through cord-cutting, and provided personalized content suggestions. OTT platforms' creation of original content has expanded the content environment. However, OTT has also raised regulatory, content licensing, and monetization issues, such as internet freedom debates and complex negotiations with content creators.

OTT has completely transformed how we access and consume entertainment in the digital age, providing viewers with increased control, convenience, and preference in their content consumption patterns.

The subject of over-the-top (OTT) analytics , which focuses on collecting, analyzing, and interpreting data produced by OTT platforms and services, is quickly developing. OTT stands for over-the-top, which describes the streaming of video, audio, and other content online instead of using conventional cable and satellite TV providers. Netflix, Amazon Prime Video, Hulu, Disney+, and other streaming services are examples of OTT platforms.

OTT Analytics is a relatively new idea that gives organisations a lot of opportunities to enhance consumer experience. It can also be used along with video analytics tools on OTT to figure out how and what kind of videos are being most popular with viewers. With this expertise, businesses may change as necessary and provide visitors with more appropriate material

Going back a few years, we can see that the growth rate of OTT platform users has been increasing up to the lockdown during the Covid-19 pandemic. This development is based on the major technology developments that allow OTT providers to provide high-quality services. the information shown on the neighbouring screens. Instead than relying only on their collection, OTT businesses invest a significant amount of money in creating content.

Due to the fast growth of OTT services and the huge quantity of data they produce, OTT analytics for study has been growing in importance in recent years. This information covers user behaviour, patterns of content consumption, device choices, geographical spread, and other things. Researchers, advertisers, creators of media, and OTT service providers can all benefit from the analysis of this data.

‘

- Here is a quick overview of the main features of OTT analytics for data study

Data gathering: As a result of user engagements with their services, OTT platforms collect an enormous amount of data. This information covers the quality of the streaming, user participation, preferred content, viewing patterns, and more. For further examination, advanced analytics systems gather this data in batch or real-time actions.

User Insights : OTT providers can better understand their audience by analysing user data. This includes user segmentation, churn rates, user retention, and demographics. These details can be used to better target certain user groups with content and marketing initiatives.

Content Performance: OTT analytics assist content producers and suppliers in evaluating the effectiveness of their work. Metrics like viewer ratings, watch time, and drop-off rates give information on what content is popular with viewers and where adjustments should be made.

Personalization: OTT providers use analytics to present users with recommendations that correspond to their needs. By analysing user data and making recommendations based on personal preferences, machine learning algorithms improve user engagement and retention.

Quality of Service (QoS): Ensuring an enjoyable user experience involves monitoring the streaming quality, including buffering, video resolution, and playback issues. OTT analytics can find and fix problems that affect QoS.

Competitive Analysis: OTT service providers utilise data analysis to compare their performance to that of their rivals in a competitive analysis. This includes user engagement analytics, content libraries, and market share statistics.

Motivation

Over-the-top (OTT) platforms' increasing power in the media and entertainment industries cannot be overstated. OTT providers have become the primary source of entertainment for a growing audience, with millions of subscribers and a vast library of content. The study was inspired by the profound impact these platforms have on our media consumption habits.

We intend to delve into the rich pool of data generated by OTT services in order to better understand user behavior and preferences, improve content recommendations, and address issues that both providers and subscribers face.

OTT platforms are changing the way we consume content, with global subscriber numbers reaching levels that were never before seen. According to industry reports, over half of all households now subscribe to at least one OTT service.

These platforms account for a significant amount of the entertainment market, making them an important focus for our research.

High subscriber churn, content recommendation accuracy, and competition are all challenges for OTT providers. These challenges, however, present opportunities for data-driven improvements, subscriber retention strategies, and overall user experience improvements.

The OTT industry is rife with competition, with many vendors vying for a piece of the pie. Analytics can help providers gain a competitive advantage by assisting them in making informed decisions, optimizing content libraries, and focusing on their offerings to user preferences.

OTT platforms, unlike traditional broadcasting, thrive on their ability to personalize content for individual users. Understanding user behavior and preferences is not only helpful, but also necessary for OTT success. The objective of this research is to elucidate the details.

Objectives

The general objective of this study is to leverage data mining and analytics to enhance user experience and content delivery in OTT platforms.

- Look into user viewing habits, like the time of day, content type, and preferred genres.
- Review user viewing patterns and preferences
- Establish popular content and viewing habits among multiple user segments.
- Develop a recommendation system that provides personalized content recommendations based on user history and preferences.
- Use user input and engagement metrics to evaluate the effectiveness of the system of suggestions.
- Figure out the factors that lead to subscriber churn, such as content dissatisfaction or pricing.
- Make recommendations for subscriber retention strategies, such as targeted content promotions and engagement campaigns.
- Ensure the privacy and security of user data in compliance with applicable regulations and standards.
- Put in place strong data security measures to protect sensitive user information.
- Assess the performance and growing popularity of existing library content.
- Look into the performance and widespread acceptance of existing library content.

LITERATURE REVIEW

In Research Paper [1],Indian OTT platforms like Hotstar and Jio Cinema compete with global giants like Netflix. The paper analyzes growth factors, technology, and audience preferences, highlighting Hotstar's dominance with non-paying users. Smartphones, mainly Xiaomi, are the primary devices for OTT consumption, and Jio leads in networking. Hindi and English are the preferred languages for viewers.

In Research Paper [2],AI-powered predictive analytics (PA) is crucial in the movie industry during COVID-19. Hollywood's response to the 'Netflix Effect' involves investing in data and AI. This review identifies key PA use cases: ScriptBook, Vault, Pilot, Cinelytic, and Merlin Video for 'compete'; Movio and Datorama for 'grow'; Industrial Light & Magic and Geena Davis Institute for 'enforce'; Watson, Benjamin, and Greenlight Essential for 'improve'; and Disney Research with Simon Fraser University for 'satisfy.'

In Research Paper [3], explores the transformative role of big data in the $103 billion global film industry, emphasizing its influence on consumer behavior, communication, and the future of cinema. With the Indian film industry valued at INR 182.2 billion, the paper delves into how big data shapes film making, investment, publicity, distribution, broadcasting, and audience engagement. It concludes that big data serves as a quantitative foundation for decision-making, contributing to revenue generation and societal impact in the film industry Top of form.

In Research Paper [18],In order to overcome conventional shortcomings, this work presents a hybrid recommendation system that combines context-based engines with collaborative filtering. It presents measurements for quality and performance and assesses recall, accuracy, and precision for content, collaborative filtering, and context filtering techniques. The hybrid engine does exceptionally well when it comes to online movie recommendations, showcasing its versatility across a variety of sectors, including music and books.

In Research Paper [7], It addresses the impact for the internet, Big Data, and Over-the-Top (OTT) TV on the entertainment industry. It clarifies the way big data analytics can be utilized to forecast consumer behavior and improve ad campaigns. The text goes further into the concept of product-to-consumer connection and the possibility of OTT in this context.

In Research Paper[33], In order to improve user models by integrating content information, this paper promotes content-based movie recommendation. The algorithm it suggests builds collaborative-based models into content-based user models, which are then adjusted based on what users want. Implicit preferences are exposed by explicit user profiles that highlight key characteristics. The system's efficacy is evaluated using decision-support metrics, and further optimization mechanisms are being investigated for better recommendations in future research.

In Research Paper [15],introduces an improved algorithm and an open-source Python library for content-based suggestions in response to the need for automated scholarly resource discovery. The algorithm performs better than keyword-based approaches when tested on papers from scientific conferences. With a focus on flexibility, instantaneous recommendations, and cooperative transparency, it offers a quick and easy way to navigate academic material.

In Research Paper [21], Using a case study of an IMDb-based movie recommendation web app, this literature highlights the pervasiveness of recommendation systems in everyday life. Using scikit-learn in Python, it prioritizes actors, director, and storyline when determining similarity scores. The study emphasizes the advantages of saving time and suggests more research on extending dataset constraints for wider consumer application.

In Research Paper[22], The Bag of Words (BOW) approach for text representation is presented in this literature, and it is used to support a Support Vector Machine classifier to build a Vector Space Model. Using publicly available datasets, the method compares raw to clean data and shows a 16.26% reduction in text data. Sentiment analysis indicates common neutral attitudes and is displayed using clouds of words and histograms.

In Research paper [25] the research looks into the dynamic changes that have occurred in television, including developments in technology and the decentralisation of content delivery. In addition to highlighting the influence on cultural citizenship and the transition from a public service ethos to a consumer-oriented approach, it emphasises the economic significance of sports content. In the midst of continuing industry transitions, the critical political economics perspective emphasises the need to protect viewer and public interests.

In Research Paper [4],Over-the-top (OTT) video services have quickly impacted conventional media distribution channels in India. The increasing number of internet-enabled mobile phones, falling data costs, and increasing access to the internet have all contributed to the growth of OTT platforms. Content usage has shifted to local languages, and sports broadcasting, particularly cricket, is becoming increasingly important. Collaborations between telecommunications companies and OTT services that offer content bundles have become common. As 5G networks become available, the industry anticipates even more possibilities for expansion. Personalization, less reliance on conventional cable TV, and a focus on regional and unique programming are changing the upcoming generations of OTT services in India.

In Research Paper [14],A variety of methods for movie recommendations are examined in the literature review. MOVREC, a collaborative filtering system for user-based suggestions, was introduced by Kumar et al. A hybrid system combining content-based and collaborative techniques was introduced by Virk et al. Kużelewska showed how clustering works well for precise recommendations, whereas De Campos suggested a collaborative approach using a Bayesian network.

In Research Paper [19*]*, The study investigates matrix factorization, user-based, and item-based techniques to collaborative filtering (CF) in recommendation systems. It highlights the combination of user similarity analysis and opinion mining for a movie recommendation system, demonstrating better outcomes than traditional methods. The suggested method outperforms ALS and SEHRS on the Movie Lens Dataset by efficiently extracting particular ratings from reviews and suggesting the top-k films.

Table 1: Comparative Studies from year 2014-2020

Sr.No	Author	Year	Dataset Used	Algorithm Used	conclusion
1	Gaurav Arora et.al [18]	2014	Not Specified	Hybrid Recommendation System (Combination of Context-based, Collaborative Filtering)	Effective hybrid engine excels in online movie recommendations, displaying versatility for music and books
2.	Steve Wong et.al [7]	2015	Not Specified	Big Data Analytics	Big Data analytics can enhance predictive buyer analytics and consumer engagement in the OTT industry.
3	KIRMEMIS et.al [33]	2015	Not Specified	Content-Based Recommendation	The paper proposes a content-based movie recommendation algorithm that enhances personalization by building user models from collaborative-based models and movie domain characteristics.
4	Titipat Achakulvisut et.al [15]	2016	Set of posters presented at the Society for Neuroscience 2015conference.	Rocchio algorithm for improving recommendations based on user votes	The system provides an open and fast approach to accurately discover new research, with potential benefits for the scientific community.
5	Sounak Bhattacharya et.al [21]	2019	IMDb-based movie data	Scikit-learn in Python	prioritizes actors, director, and storyline for similarity scores. Emphasizes time-saving benefits. Recommends further research for

					expanding dataset constraints for broader consumer application.
6	Dr. Ashok Kumar R et.al [22]	2019	Publicly available datasets	Bag of Words (BOW)	Presents BOW for text representation, supports SVM classifier to build Vector Space Model. Compares raw to clean data, showing a 16.26% reduction in text data. Sentiment analysis reveals common neutral attitudes through word clouds and histograms.
7	Brwhett Hutchins et.al [25]	2019	Not Specified	Not Specified	it focuses on analyzing the growth of over-the-top (OTT) Internet and mobile video streaming services in the context of global media sports distribution.
8	E. Sundaravel et.al[4]	2020	Not Specified	Not Specified	OTT platforms are growing in India due to increased mobile internet access and a shift towards regional content.

9	F. Furtado et.al[14]	2020	Movies Dataset	Matrix Decomposition, Clustering, Deep Learning	The recommendation system is collaborative-based, offering explicit outcomes compared to content-based systems.
10	Yogesh Naveen Kumar et.al[19]	2020	Movie Lens Dataset	Opinion Mining and User Similarity Analysis	Hybrid system recommends top-k movies successfully, outperforming ALS and SEHRS, enhancing accuracy and meeting user requirements.

After studying above literature reviews we conclude that, The literatures reveals a dynamic entertainment landscape shaped by global OTT platform competition, notably Hotstar and Jio Cinema in India, emphasizing smartphone prevalence and language preferences. Hollywood's response to the 'Netflix Effect' involves substantial investments in AI and data analytics. Big data emerges as a transformative force in the $103 billion global film industry, influencing filmmaking, investment, and audience engagement. Innovative approaches, such as hybrid recommendation systems and content-based movie recommendations, highlight the industry's commitment to enhancing user experiences. The literature collectively underscores the industry's evolution at the intersection of technology, data analytics, and changing consumer behaviors.

In Research Paper [8],The article discusses the increasing presence of Indian customers on over-the-top (OTT) platforms, owing to improved networks and mobile phones. It emphasizes Disney+ Hotstar, Amazon Prime Video, and Netflix's dominant position in the Indian OTT marketplace, as well as the increase in demand throughout the COVID-19 pandemic. The research additionally highlights Indian viewers' preference for native language content. It presents the results of a survey on OTT platform usage, membership costs, and user satisfaction. According to the study, India's OTT industry will grow significantly in the years to come.

In Research Paper [16],The sense of ownership of over-the-top (OTT) services and efficient content recommendation are the subjects of this study. While users with little ownership respond better to abstract messaging, individuals with strong ownership prefer objective movie information. The study validates the function of psychological distance as a mediator. Theoretical implications point to the use of ownership to customise tactics and accurate content.

In Research Paper [17],The current research analyses tag genes, age, and genre as determinants of OTT subscription preferences. When it comes to TV-MA films, Netflix is the leader, but Amazon Prime has a more evenly distributed maturity rating. The family, adventure, and animat0ion genres are where Disney+ shines. According to genre study, Netflix has the widest selection. Genre findings are consistent with genome-tag analytics. One of the limitations is the smaller dataset for original films, which means that additional analysis with new releases is required to ensure accuracy.

In Research Paper [20], The body of research highlights how important movie recommendation engines are for reducing user effort while searching for material. It criticizes current approaches, in particular Content-Based Filtering, and suggests a novel hybrid strategy that makes use of several text-to-vector conversion techniques. The summary states that better recommendations are obtained by manipulating the algorithms, demonstrating the benefits of using separate algorithms and bolstering the efficacy of the hybrid approach.

In Research paper [26] The importance of recommendation algorithms is emphasised in this evaluation, especially with regard to streaming video services. It suggests a machine learning method for better movie rating prediction and recommendation that makes use of stacked autoencoders and Singular Value Decomposition. The system performs better than conventional algorithms, demonstrating how autoencoders may reduce dimensionality and improve accuracy.

In Research paper [28] In the OTT sector, this evaluation contrasts Netflix with Amazon Prime Video (APV). Whereas APV suffers from an unpolished user interface, Netflix thrives at providing personalised service and a straightforward subscription plan. The report highlights the changing competitive dynamics in the OTT landscape and highlights Netflix's success in content selection and teamwork.

In Research Paper [29], This paper presents two well-known recommender systems: a context-based recommender system that makes suggestions based on previous interactions and a context-aware recommender system that takes user interests into account. It delivers personalized news, filters e-learning content, and supports media recommendations for smartphones through the use of content-based, collaborative filtering and hybrid approaches. Using E-paper and semantic reasoning, it works like a content-based recommender system, presenting options based on user ratings, interests, and evaluations.

In Research Paper [23], The literature emphasizes how information gathering, learning techniques, IoT, and deep learning are driving the increased importance of recommendation systems, especially content-based algorithms. In the face of copious amounts of data, it highlights accuracy in content-based recommender systems and presents a novel technique for enhancing movie representation. The suggested framework combines ongoing data, especially in the medical field, to offer customized and pertinent support.

In Research paper [24] The difficulties of browsing large movie datasets on online platforms are examined in this review of the literature. The study explores movie recommender systems and uses Tableau visualisation and exploratory data analysis to identify the elements driving a movie's appeal. Important factors include historical performance, actor and director recognition, genre, and geographic location. The study provides auto-suggestions and thorough information retrieval on searched movies, emphasising data cleaning, preprocessing, and model training for efficient user engagement.

In Research paper [27] The article discusses the rising need for improved database administration in light of the growing popularity of over-the-top platforms. It draws attention to the rivalry between OTT service providers over user experience and content quality. In order to improve user experience, the report advises improving data management by providing comprehensive search and sorting options based on many criteria. To improve OTT platforms, the emphasis is on reducing data redundancy and adjusting to modern needs, particularly those that arise after COVID-19.

In Research Paper [30], This recommendation system utilizes content-based filtering to suggest movies based on similar qualities, with the goal of predicting user preferences for movies. Ten movie recommendations are given to users, along with information about the films' cast, rating, release date, and genres. To help users make decisions, the system also analyzes reviews sentiment and classifies them as "Good" or "Bad."

In Research Paper [5],The research examines customer preferences in the OTT platform nature and presents a recommendation system based on the Random Forest algorithm. The system intends to offer specific platform suggestions through analyzing data from 410 subjects, thereby increasing customer satisfaction and engagement. Other methods were outperformed by the Random Forest algorithm, which achieved greater accuracy, recollection, and F1-scores. This study helps to improve the user experience when choosing the best OTT platform and provides valuable industry insights.

In Research Paper [6],Dr.Rajendra Prasad Meena And Suman Kumar's study looks at changing consumer preferences toward Over-The-Top (OTT) platforms in India. The popularity of platforms such as YouTube and Netflix, the effect of social media and closely suggestions, and the effects of COVID-19 lockdown on grew OTT consumption are among the findings. Furthermore, the research reveals that consumers prioritize factors such as cost-effectiveness, content quality, and limited advertising. Despite some limitations, the study emphasizes the growth potential of OTT platforms in India, as well as the importance of quality content and focused on customers strategies in order to succeed in this highly competitive sector.

In Research Paper [9],The Article Explains The COVID-19 pandemic changed the face of entertainment, with OTT platforms gaining traction as traditional theaters faded. The rise in OTT popularity, fueled by simplicity and appealing material, calls into question theater's distinct aesthetics and the audience appeal. Despite that, Indian theater can adapt as well as compete by drawing on a long tradition of theater gratitude. Both mediums coexist, resulting in a developing entertainment scenery in which theater faces new challenges while retaining its essence.

In Research Paper[10], Due to theater closures and restrictions, the COVID-19 pandemic caused an increase in Over-The-Top (OTT) platforms in India. OTT acceptance became common during this time period, providing families with convenience and cost savings. Users are satisfied with their subscription, indicating a promising future for the OTT industry as the web and mobile penetration grows, shaping new customer tastes and viewing habits.

In Research Paper[11],The paper discusses framework types and algorithms while outlining the concept and architecture of a Live OTT platform. It looks into the difficulties encountered, such as gaining trust and collaborating with theaters in the Indian market. As seen with platforms like Netflix, successful collaborations can transform the OTT industry. Complex layouts and resource management are critical for such systems. Algorithms play an important role, but they must be updated and optimized on a regular basis for the best performance and video streaming efficiency

In Research Paper [12],The Article Discusses The introduction of OTT services has transformed media consumption, especially in India, where the young and technologically savvy population seeks different kinds of media. OTT subscriptions increased during the COVID-19 pandemic as people sought alternative entertainment at home. The ease of use, variety of content, and reasonable cost of OTT are key drivers of adoption. While mobile accessibility and personalization are appealing to users, the big-screen expertise is sacrificed. Netflix is a popular platform, but it lacks the cinematic atmosphere of traditional theaters.

In Research Paper [13],The paper’s study looks into why people trust and continue to use a subscription-based streaming services like Netflix. They employed a model that considers various factors such as the value people derive from these platforms and their level of trust in them. According to the study, specific variables such as the value of the content and how many individuals trust the platform impact whether they continue to use it. This study is significant for the online entertainment industry because it explains why people continue to use services like Netflix.

In Research Paper[31], The research proposes a Movie Recommendation System that maps movie vectors using vectorization and computes similarities based on angles, but it is item-based filtering instead of user-based. It makes use of feature engineering, data preprocessing, and models such as content-based and item-based collaborative filtering. The study emphasizes how successful this method is when compared to user-based collaborative filtering, which is employed by well-known services like Netflix and Amazon Prime.

In Research Paper[s], To make users' movie selections easier, the Movie Recommendation system uses cosine similarity, count vectorization, and machine learning algorithms. It offers URLs for in-depth descriptions and suggests the top 5 films based on user interests, addressing the challenge of sifting through millions of films. The goal of this project is to use cutting-edge technologies to improve the movie-watching experience for users.

Table 2: Comparative Studies from year 2021 – 2023

Sr.No	Author	Year	Dataset Used	Algorithm Used	conclusion
1	Bhavyarajsinh D. et.al[8]	2021	Survey data from Indian consumers	Not Specified	The study predicts significant growth in India's OTT industry, driven by consumer behavior shifts during the COVID-19 pandemic.
2	Bong-Goon Seo et.al[16]	2021	Not Specified	Not Specified	Study emphasizes psychological ownership in online services, influencing content decisions
3	Rick Kim et.al[17]	2021	Kaggle datasets	Not Specified	Netflix dominates TV-MA; Disney+ excels in family, adventure, animation. Genre analytics align. Further analysis, recent movies needed.
4	Rahul Pradhan et.al[20]	2021	Not Specified	Collaborative Filtering, Hybrid Movie Recommendation System	Collaborative Filtering is powerful but faces runtime and data sparsity issues. Hybrid system proposed to address limitations. Future work includes addressing weaknesses and improving the user interface.
5	R. Lavanya et.al[26]	2021	IMDb-based movie data	Scikit-learn in Python	rioritizes actors, director, and storyline for similarity scores. Emphasizes time-saving benefits. Recommends further research for expanding dataset constraints for broader consumer application.

6	Hansei University, Korea et.al[28]	2021	Movie Lens dataset	Auto Encoders, Singular Value Decomposition	The paper proposes a recommendation system for streaming movie services using machine learning, specifically based on auto encoders for collaborative filtering. The Movie Lens dataset is utilized
7	Umair Javed et.al[29]	2021	Not Specified	Content-Based, Collaborative Filtering, Hybrid Recommender System,	Review explores context-aware, context-based recommender systems using diverse algorithms
8	Dr. Ganesh et.al[23]	2022	Movie dataset, Feature extraction, Cosine	Content-Based Filtering Algorithm	Implements a content-based filtering algorithm with feature extraction. Transforms relevant features into vectors and determines similarity using the cosine similarity algorithm. Generates movie recommendations based on the similarity of feature vectors.
9	Heet Shah et.al[24]	2022	Kaggle and IMDB	Exploratory Data Analysis	The analysis involved cleaning the data, extracting valuable insights, handling missing values, and preprocessing the data for model training.
10	A. Shah et.al[27]	2022	Not Specified	Not Specified	it emphasizes the rising need for improved database administration in the context of the growing popularity of over-the-top (OTT) platforms.

11	Shubham Pawar et.al[30]	2022	Not Specified	Content-Based Filtering	The recommendation system utilizes content-based filtering to suggest movies based on similar qualities, aiming to predict user preferences.
12	Md Samiul Islam et.al[5]	2023	Data from 410 subjects	Random Forest Algorithm	The Random Forest algorithm outperforms other methods for recommending OTT platforms, improving user satisfaction
13	Suman Kumar et.al[6]	2023	Not Specified	Not Specified	Changing consumer preferences in India favor OTT platforms with quality content and cost-effectiveness
14	Prof.Durgesh Ravande et.al[9]	2023	Not Specified	Not Specified	COVID-19 pandemic reshaped entertainment, propelling OTT platforms while challenging theater's aesthetics and appeal.
15	Mythili .D et.al[10]	2023	Not Specified	Not Specified	Increased OTT usage during the pandemic; users satisfied, promising future for the industry.
16	Dr.Shivi et.al[11]	2023	Secondary research papers, internet articles	Collaborative filtering, Content-Based algorithms	This innovative concept may face challenges in theater partnerships but has the potential to revolutionize OTT, enriching home cinema experiences
17	T.Manoj Kumar et.al[12]	2023	Not Specified	Not Specified	The Ease of use, content variety, and reasonable cost are key drivers of OTT adoption.

18	Debarun Chakraborty et.al[13]	2023	Survey Data	Not Specified	Trust and intent to purchase in OTT platforms are influenced by identified values, with trust acting as a mediator. Knowing these values is critical for the success of an OTT platform.
19	Aayush Khanna et.al[31]	2023	Not Specified	Item-Based Filtering, Vectorization	he Movie Recommendation System utilizes item-based filtering, employing vectorization to map movies onto a single point and determine similarities based on angles between vectors.
20	Shreenivas Satish Kulkarni et.al[32]	2023	Not Specified	Cosine Similarity, Count Vectorization, Machine Learning Algorithms	he Movie Recommendation system employs cosine similarity, count vectorization, and machine learning algorithms to simplify users' movies.

From the above literature reviews we conclude that,The comprehensive analysis anticipates substantial growth in India's OTT industry, spurred by shifts in consumer behavior amid the COVID-19 pandemic. It underscores the psychological ownership in online services influencing content decisions, with Netflix dominating TV-MA content and Disney+ excelling in specific genres. The study advocates for a hybrid recommendation system to overcome collaborative filtering limitations, prioritizing actors, directors, and storylines. Additionally, it proposes a machine learning-based recommendation system using autoencoders on the Movie Lens dataset. The exploration of context-aware recommender systems, content-based filtering algorithms, and the utilization of the Random Forest algorithm for OTT platform recommendations further enriches the understanding of evolving consumer preferences and the industry's dynamic landscape.

Dataset Preparation

We Have Used This Dataset For Our Project.This Dataset Is Taken From Kaggle.Com

https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata

Dataset 1: tmdb_5000_movies:-

- budget:The budget allocated for the production of the movie or TV show.
- genres: The categories or genres to which the content belongs (e.g., action, drama, comedy).
- homepage: The official website or online platform associated with the movie or TV show.
- id: A unique identifier for each entry in the dataset.
- keywords:Keywords or phrases associated with the content, indicating its themes or topics.
- original_language:The language in which the content was originally produced.
- original_title:The original title of the movie or TV show in its source language.
- overview: A brief summary or description of the content.
- popularity:A measure of the content's popularity, likely based on user engagement or views.
- production_companies: The companies involved in the production of the content.
- production_countries: The countries where the content was produced.
- release_date: The date when the content was released.
- revenue: The earnings or revenue generated by the content.
- runtime: The duration of the movie or TV show in terms of running time.
- spoken_languages: The languages spoken in the content.
- status: The current status of the content (e.g., released, in production).
- tagline: A catchy phrase or slogan associated with the content.
- title: The title of the movie or TV show.
- vote_average: The average rating given by users or viewers.

- vote_count: The number of votes or ratings received by the content.

Dataset 2:tmdb_5000_credits:-

- movie_id: A unique identifier for each movie in the dataset.
- title: The title of the movie.
- cast: Information about the cast members involved in the movie.
- crew: Information about the crew members involved in the movie. This could include directors, producers, writers, etc.

Data Preprocessing

I started my project with data collection and pre-processing. In order to achieve best prediction model, we must pre-process the dataset. The following steps are performed In My Project:

1.Data Loading:

You load two datasets: 'tmdb_5000_movies.csv' and 'tmdb_5000_credits.csv' using pd.read_csv.

2.Data Merging:

You merge the two datasets based on the 'title' column using the merge function.

3.Data Cleaning:

You select specific columns of interest ('movie_id', 'title', 'overview', 'genres', 'keywords', 'cast', 'crew').

You check for missing values and drop rows with missing data.

You check for and drop duplicate records.

4.Feature Engineering:

You extract relevant information from the 'genres', 'keywords', 'cast', and 'crew' columns.

You define functions to convert string representations of lists to actual lists using the ast module.

You limit the number of cast members to three in the 'convert3' function.

You extract the director from the 'crew' column using the 'fetch_director' function.

You tokenize the 'overview' column by splitting it into words.

5.Text Processing and Tag Creation:

You process and clean the 'genres', 'keywords', 'cast', 'crew', and 'overview' columns.

You create a new column 'tags' by combining information from 'overview', 'genres', 'keywords', 'cast', and 'crew'.

6.Manipulation:

You create a new DataFrame ('new_df') with columns 'movie_id', 'title', and 'tags'.

You concatenate the cleaned and processed tags into a single string for each movie.

7.Text Normalization:

You use stemming to normalize the text in the 'tags' column.

8.Vectorization:

You use the CountVectorizer from scikit-learn to convert the 'tags' into a numerical vector representation.

9.Cosine Similarity Calculation:

You calculate cosine similarity between the vectors representing different movies.

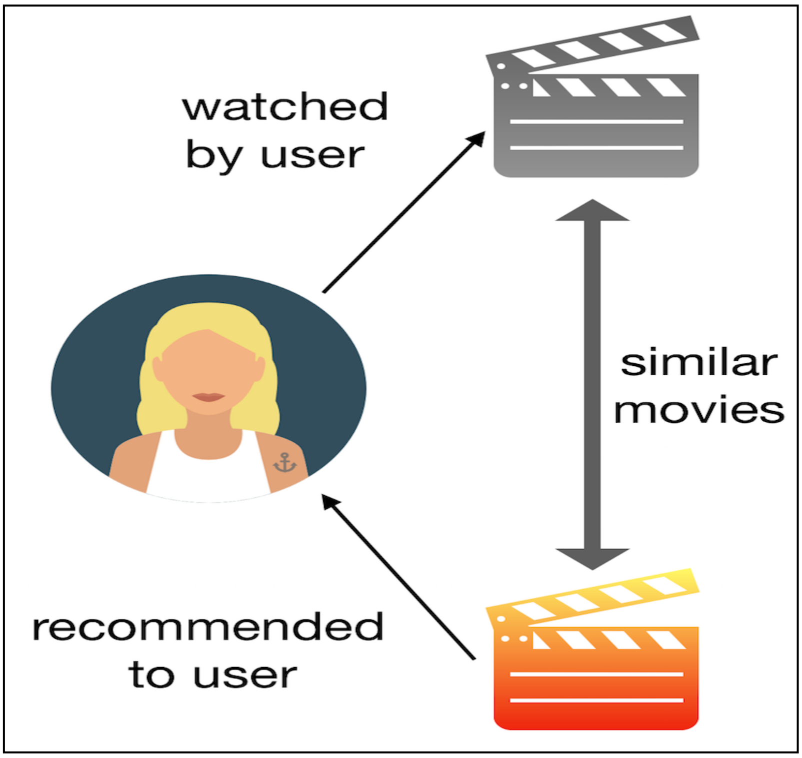
10.Movie Recommendation Function:

You define a function ('recommend') that takes a movie title as input and recommends similar movies based on cosine similarity.

11.Model Storage:

You use the pickle module to save the processed DataFrame ('new_df'), a dictionary version of it, and the cosine similarity matrix.

<u>Content-Based Recommendation System Implementation</u>



It focuses on data preprocessing, feature engineering, and building a content-based recommendation system using cosine similarity. The core algorithm is based on comparing the similarity between movies using their textual features.

Content-Based Filtering:

- Algorithm Description: Content-based filtering recommends items similar to those the user has shown interest in based on item features.
- Usage in Code: The recommendation is made by calculating the cosine similarity between movie descriptions, genres, keywords, cast, and crew.

Text Processing and Tag Creation:

- Algorithm Description: Tokenization, cleaning, and merging of textual features (overview, genres, keywords, cast, crew) to create a consolidated set of tags representing each movie.
- Usage in Code: Utilized to extract meaningful information from text and represent it in a standardized format.

```
#merging overview,genres,keywords,cast,crew to Tag
movies['genres'] = movies['genres'].apply(lambda x:[i.replace(" ","") for i in x])
movies['keywords'] = movies['keywords'].apply(lambda x:[i.replace(" ","") for i in x])
movies['cast'] = movies['cast'].apply(lambda x:[i.replace(" ","") for i in x])
movies['crew'] = movies['crew'].apply(lambda x:[i.replace(" ","") for i in x])
```

```
movies.head()
```

```
movies['tags']=movies['overview']+movies['genres']+movies['keywords']+movies['crew']+movies['cast']
```

Stemming:

- Algorithm Description: Reducing words to their root or base form to capture their essential meaning.
- Usage in Code: Applied to normalize text in the 'tags' column, reducing words to their stems.

```python
from nltk.stem.porter import PorterStemmer
ps = PorterStemmer()
```

```python
def stem(text):
    y = []

    for i in text.split():
        y.append(ps.stem(i))

    return " ".join(y)
```

Vectorization (CountVectorizer):

- Algorithm Description: Converts text data into numerical vectors, representing the frequency of words in the text.
- Usage in Code: The CountVectorizer from scikit-learn is used to convert the 'tags' into a matrix of token counts.

```python
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer(max_features = 5000, stop_words='english')
```

```python
vectors = cv.fit_transform(new_df['tags']).toarray()
```

```python
vectors
```

```
array([[0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       ...,
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0]], dtype=int64)
```

Cosine Similarity:

- Algorithm Description: Measures the cosine of the angle between two non-zero vectors, providing a measure of similarity.
- Usage in Code: Cosine similarity is calculated between vectors representing different movies to determine their similarity.

```
from sklearn.metrics.pairwise import cosine_similarity
```

```
similarity = cosine_similarity(vectors)
```

```
similarity[2]
```

```
array([0.0860309 , 0.06063391, 1.        , ..., 0.02451452, 0.        ,
       0.        ])
```

```
sorted(list(enumerate(similarity[0])), reverse = True, key=lambda x:x[1])[1:6]
```

```
[(1216, 0.2867696673382022),
 (2409, 0.26901379342448517),
 (3730, 0.2605130246476754),
 (507, 0.255608593705383),
 (539, 0.25038669783359574)]
```

Recommendation Function:

- Algorithm Description: Sorting and recommending movies based on their similarity scores.
- Usage in Code: The recommend function takes a movie title as input, finds its index, calculates similarity scores, and recommends similar movies.

```
def recommend(movie):
    movie_index = new_df[new_df['title']==movie].index[0]
    distances = similarity[movie_index]
    movies_list = sorted(list(enumerate(distances)), reverse = True, key=lambda x:x[1])[1:6]

    for i in movies_list:
        print(new_df.iloc[i[0]].title)

    return
```

Results:

```
recommend('Avatar')
```

```
Aliens vs Predator: Requiem
Aliens
Falcon Rising
Independence Day
Titan A.E.
```

```
recommend('Batman Begins')
```

```
The Dark Knight
Batman
Batman
The Dark Knight Rises
10th & Wolf
```

```
recommend('The Dark Knight Rises')
```

```
The Dark Knight
Batman Returns
Batman
Batman Forever
Batman Begins
```

Conclusion

In conclusion, the implemented content-based movie recommendation system demonstrates a practical application of leveraging textual features for personalized suggestions. By extracting information from movie descriptions, genres, keywords, cast, and crew, the system creates cohesive tags that capture the essence of each film. The use of cosine similarity effectively measures the likeness between movies, forming the basis for relevant recommendations.

This content-based approach excels in providing user-specific suggestions, aligning with individual tastes and preferences. The recommendation function, powered by cosine similarity scores, offers a transparent and interpretable way to suggest movies based on textual content. The model storage using the pickle module ensures the persistence and reusability of the recommendation system.

Looking forward, potential enhancements could involve exploring additional textual features or integrating collaborative filtering techniques for a more comprehensive recommendation strategy. Despite its current focus on content, the system lays a foundation for future iterations and improvements in personalized movie recommendations. Overall, the project successfully addresses the need for tailored suggestions based on movie content, contributing to a user-friendly and adaptable recommendation system.

References

[1]Sundaravel, E. and Elangovan, N. (2020). Emergence and future of over-the-top (OTT) video services in India: an analytical research. International Journal of Business, Management and Social Research, 08(02), 489-499.https://doi.org/10.18801/ijbmsr.080220.50

[2]M. Song, "A Study on the Predictive Analytics Powered by the Artificial Intelligence in the Movie Industry," International journal of advanced smart convergence, vol. 10, no. 4, pp. 72–83, Dec. 2021.

[3]Sengar, P. (2020). Big Data: Frame the Cinema from Viewership to Projection. MediaSpace: DME Media Journal of Communication, 1(01), 66–72. https://doi.org/10.53361/dmejc.v1i02.8

[4]Sundaravel, E. and Elangovan, N. (2020). Emergence and future of over-the-top (OTT) video services in India: an analytical research. International Journal of Business, Management and Social Research, 08(02), 489-499.https://doi.org/10.18801/ijbmsr.080220.50

[5]Md Samiul Islam*1, Pierre Mukisa Tshomba*2, Songjiang Li*3 *1Department of Computer Science and Technology, Changchun University of Science and Technology, Jilin, *3 Professor, Department of Computer Science and Technology, Changchun University of Science and Technology, Jilin, China DOI : https://www.doi.org/10.56726/IRJMETS41702

[6]Kumar, Suman & Meena, Rajendra & Scholar, Research. (2023). The Rise of OTT Platform: Changing Consumer Preferences. INTERNATIONAL JOURNAL OF MANAGEMENT. 15. 70-94. 10.17605/OSF.IO/AN69G.

[7]Wong, Steve. (2015). Over The Top TV (OTT), Big Data and The Future Of Predictive Buyer Analytics.

[8]Jhala, Bhavyarajsinh & Patadiya, Vivek. (2021). A STUDY ON CONSUMER BEHAVIOUR TOWARDS OTT PLATFORMS IN INDIA DURING COVID ERA. 08. 120-125.

[9]Durgesh Bhausaheb Ravande, & Gulab Karim Shaikh. (2023). OTT Platform and Dynamics for Contemporary Indian Theatre. *Creative Saplings*, *2*(06), 36–47. https://doi.org/10.56062/gtrs.2023.2.06.381

[10].D, Mythili. (2023). AN ANALYSIS OF CUSTOMER PERCEPTION ON OTT. VIII. 59-66.

[11]Sharan S, Shri & Sharma, Shivi & Reddy, Dinesh & Fayaz, Mohammad & Karthikeyya, Krovvidi & Sravan, Baddireddy. (2023). Live OTT Platform.

[12]Thamotharan, Manoj & L., Ivan Kenny Raj. (2023). A Study On Preference Of College Students Towards Ott Platform. 21. 72-80.

[13]Debarun Chakraborty, Mujahid Siddiqui, Aaliyah Siddiqui, Justin Paul, Ganesh Dash, Francesca Dal Mas,Watching is valuable: Consumer views – Content consumption on OTT platforms,Journal of Retailing and Consumer Services,Volume 70,2023,103148,ISSN 0969-6989,https://doi.org/10.1016/j.jretconser.2022.103148.

[14]Giridharan, N. & Nathan, K. & Swetha, M.. (2022). Movie recommendation system using machine learning. International journal of health sciences. 498-505. 10.53730/ijhs.v6nS8.9728.

[15]Achakulvisut T, Acuna DE, Ruangrong T, Kording K (2016) Science Concierge: A Fast Content-Based Recommendation System for Scientific Publications. PLoS ONE 11(7): e0158423. doi:10.1371/journal.pone.0158423

[16]Seo, B.-G.; Park, D.-H. Effective Strategies for Contents Recommendation Based on Psychological Ownership of over the Top Services in Cyberspace. *J. Theor. Appl. Electron. Commer. Res.* 2021, *16*, 976-991. https://doi.org/10.3390/jtaer16040055

[17**]**. Data Analysis on OTT Platforms**.**Rik Kim, Published in Towards Datascience,may 20 2021- https://towardsdatascience.com/data-analysis-on-ott-platforms-which-service-should-i-choose-8eed953ff7d2

[18] Gaurav Arora1 , Ashish Kumar2 , Gitanjali Sanjay Devre3 , Prof. Amit Ghumare4 1,2,3Student, Computer Engineering Department 4Assistant Professor, Computer Engineering Department 1,2,3,4Dr. D. Y. Patil College of Engineering, Ambi University of Pune.https://www.academia.edu/6867161/MOVIE_RECOMMENDATION_SYSTEM_BASED_ON_USERS_SIMILARITY_

[19]Kumar, Yogesh & Kumari, Naveen. (2020). Movie Recommendation System.- https://www.researchgate.net/publication/344633593_Movie_Recommendation_System

[20]Rahul Pradhan *et al* 2021 *IOP Conf. Ser.: Mater. Sci. Eng.* 1119 012018DOI 10.1088/1757-899X/1119/1/012018

[21]Sounak Bhattacharya Ankit Lundia Department of CSE Department of CSE SRM Institute, Ankit Lundia Department of CSE Department of CSE International Journal of Engineering Applied Sciences and Technology, 2019 Vol. 4, Issue 5, ISSN No. 2455-2143- https://typeset.io/pdf/movie-recommendation-system-using-bag-of-words-and-scikit-4v0uxdvc9u.pdf

[22]Chingmuankim , Rajni Jindal Classification and Analysis of Textual data using Naïve Bayes with TF-IDF
https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID3661173_code3253098.pdf

[23]Dr. Ganesh D Associate Professor, School of CS & IT, Yash Bhansali , Department of MCA, School of CS & IThttps://ijcrt.org/papers/IJCRT2204045.pdf

24.Platforms With Data Analytics Techniques For Recommending Personalized Content To The Users- Heet Shah, Vaishnavi Jadhav, Tanmay Gharte, Soham Wattamwar, Varsha Naik

[24]Dr. Vishwanath Karad MIT World Peace University Heet Shah, Evaluation Of Different Ott Platforms With Data Analytics Techniques For Recommending Personalized Content To The Users.https://www.ijrar.org/papers/IJRAR1CNP012.pdf

[25]Hutchins, B., Li, B., & Rowe, D. (2019). Over-the-top sport: live streaming services, changing coverage rights markets and the growth of media sport portals. Media, Culture & Society, 41(7), 975-994. https://doi.org/10.1177/0163443719857623

[26]R. Lavanya Assistant Professor, Ebani Gogia UG Scholar, Department of Computer Science and Engineering, DOI:10.14704/WEB/V18SI04/WEB18285

[27]Aryan Shah, Charmi Shah-Proposed DBMS for OTT platforms in line with new age requirements https://arxiv.org/ftp/arxiv/papers/2106/2106.00932.pdf

[28]International Journal of Advanced Smart Convergence Vol.10 No.1 62-74 (2021) http://dx.doi.org/10.7236/IJASC.2021.10.1.62

[29]Javed, U., Shaukat, K., A. Hameed, I., Iqbal, F., Mahboob Alam, T., & Luo, S. (2021). A Review of Content-Based and Context-Based Recommendation Systems. International Journal of Emerging Technologies in Learning (iJET), 16(03), pp. 274–306. https://doi.org/10.3991/ijet.v16i03.18851

[30]Shubham Pawar, Pritesh Patne, Priya Ratanghayra, Simran Dadhich, Shree Jaswal IT dept., SFIT,Mumbai,India,https://www.academia.edu/download/99915000/IJISRT22APR1053_201.pdf

[31]Aayush Khanna, Computer Science Engineering, Kartik Computer Science Engineering- Movie Recommendation System Using Machine Learning-

https://easychair.org/publications/preprint_download/pZQS

[32]Shreenivas Satish Kulkarni*1, Dr. M. Lakshmi*2 *1,2Computer Science & Engineering, SRM Institute of Science and Technology Chennai , India. DOI : https://www.doi.org/10.56726/IRJMETS38384

[33]Department of Computer Engineering, METU, Ankara, TURKEY 0090 312 210 20 80 {e120321,birturk}@metu.edu.tr.https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=fd7485eb4545f36d00664088c7bccce422e69ba2