International Journal of Creative and Open Research in Engineering and Management
ISSN: 3108-1754 (Online)
Volume 02 Issue 10 October-2026 | Impact Factor: 3.5



# OTT Content Analytics and Predictive Machine Learning: A Structured Review of Trends, Forecasting, and Recommendation Methods

**Roshani D. Kolhe[1], Dr. Sushma V. Telrande[2], Assistant Prof. Priyadarshani S. Gotmare[3]**
[1] Department of Computer Science & Engineering, Guru Nanak Institute of Engineering and Technology, Nagpur, India
[2] (Head of Department),Department of Computer Science & Engineering, Guru Nanak Institute of Engineering and Technology, Nagpur, India
[3](Assistant Professor), Department of Computer Science & Engineering,

Guru Nanak Institute of Engineering and Technology, Nagpur, India

Corresponding Author Email: kolhenikita23@gmail.com

## Abstract—

The fast growth of over-the-top (OTT) media streaming has greatly changed the way films and television shows are made and enjoyed. The growing availability of metadata from digital content allows one to use data analysis and machine learning for exploring the trends in the content, understanding its popularity, and making decisions based on facts. The paper discusses the research on OTT content analysis, prediction of its popularity, time-series forecasting, machine learning algorithms, and recommender systems. The limitations of the traditional machine learning methods, trends analysis methods and deep learning technology are reviewed and analyzed. Special attention is given to the problems related to the quality of data, feature engineering, changes in the preferences of the audience, and limitations of the usage of popularity indicators of the content instead of information related to audience viewing. A new analytical framework is proposed based on the information studied for the analysis of films and television metadata collected from The Movie Database (TMDb) between 2020 and 2025. The review also emphasizes the need to address research shortages related to the lack of uniformity in ways of assessing popularity, missing databases, issues of time sensitivity, lack of information about audience interaction, and model generalizability. The framework introduced is a starting point for systematic OTT content analysis and predictions. Its predictive power has to be demonstrated in actual studies rather than assumed off the bat.

**Keywords—** OTT Platforms, Content Analytics, Predictive Machine Learning, Popularity Prediction, Time-Series Forecasting, Recommender Systems, TMDb.

**How to Cite this Article**:

Kolhe, R. D., Telrande, S. V. & Gotmare, A. P. P. S. (2026). OTT Content Analytics and Predictive Machine Learning: A Structured Review of Trends, Forecasting, and Recommendation Methods. International Journal of Creative and Open Research in Engineering and Management, 02(10), 1-9. https://doi.org/10.55041/ijcope.v2i10.215

**License:**

This article is published under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

© The Author(s). Published by International Journal of Creative and Open Research in Engineering and Management.







https://doi.org/10.55041/ijcope.v2i10.215

## I. INTRODUCTION

The swift rise of Over-the-Top (OTT) streaming services has changed the way movies and television shows are provided, received, and viewed. These kinds of services contain a variety of content in several languages, kinds, release periods, and audiences. As the amount of content keeps increasing, data analytics and machine learning are being applied to investigate the properties of content.

OTT content analytics refers to the procedures involving the collecting, processing and scrutinising of information about films and videos. During such processes, metadata about the films including release dates, genres, original languages, popularity scores and audience ratings may be analysed in order



identify the ways of content distribution and find the evolution trends. However, missing values, inconsistent metadata and differences in the nature of content may affect the accuracy of data analysis. For this reason, effective data preprocessing, data exploratory analysis and feature extraction become integral to the development of a proper content analytics framework.

Research has been conducted into the use of machine learning to predict the success of new video content published online. For example, Jeon et al. suggested a combined approach using gradient-boosting methods and neural network technology to predict the success of various pieces of content [1]. Furthermore, extensive use of recommendation systems has been made for implementing user-oriented content search. For instance, a paper by Gomez-Uribe and Hunt reviewed the Netflix recommendation system from the perspective of its algorithms, value for business, and evaluation processes [2]. Lastly, there are new surveys on the application of recommendation techniques based on deep learning [3], whereas other researchers focused their attention on deep reinforcement learning methods [4].

While popularity prediction and personalized recommendation are distinct fields of research, they serve different aims. Popularity prediction measures the amount of interest in a particular content, while recommendation systems help recognize content relevant to specific users. The study of both fields combined with explorative data analysis and forecasting techniques lays down a solid groundwork for delineating the trends in the OTT content industry.

The present review paper investigates the relevant literature on OTT content analytics, machine learning-based popularity prediction systems, time series forecasting, and recommendations. The review highlights approaches adopted and their possible applications, evaluation aspects, and limitations of some methods. Special attention is paid to the quality of the data available, changes in content popularity over time, audience-related information, and the difficulties of the comparison of different prediction methods.

After going through the literature, the paper presents its proposed framework for analyzing and predicting OTT content. The framework comprises collecting the data, preprocessing, exploratory data analysis, applying feature engineering, prediction, evaluation, predicting the trends, and visualizing the findings. The aim of developing this framework is to organize the systematized analysis of the metadata of films and television content and to make way for the experimental applications later on.

## II. REVIEW METHODOLOGY

After going through the literature, the paper presents its proposed framework for analysing and predicting OTT content. The framework comprises collecting the data, preprocessing, exploratory data analysis, applying feature engineering, prediction, evaluation, predicting the trends, and visualizing the findings. The aim of developing this framework is to organize the systematized analysis of the metadata of films and television content and to make way for the experimental applications later on.

### A. Review Objectives

This review will focus on the following four objectives:

1. Examining the analytical methods for studying the features of film and television content.
2. Looking into machine learning and time-series methods for predicting the popularity of content and predicting trends.
3. Analysing recommendation methods including traditional and deep learning methods.
4. Identifying existing research gaps in terms of data quality, time changes, evaluation methods, and the availability of audience-level information.

### B. Literature Search Strategy

The relevant publications are obtained through scholarly search engines, such as Google Scholar, IEEE Xplore, ACM Digital Library, ScienceDirect, and Springer Link. Keywords are being used as part of the search process. These keywords reflect the focus of the review, such as “OTT content analytics,” “online video popularity prediction,” “machine learning for content forecasting,” “time-series forecasting,” “deep learning recommender systems,” and “streaming platform recommendation algorithms.”



The use of Boolean combinations can help narrow down the searches to find the papers about their objectives. The search is conducted around the materials that discuss methodologies of analysis, methods of prediction, methods of recommendation, evaluation of the methods, or limitations of OTT. The selected studies cover both classic and modern works that help understand machine learning and its applications in recommender systems.

### C. Study Selection Criteria

The relevance of studies is determined on the basis of their coverage of some of the areas that allow concluding about at least one of the following: predicting content popularity, using machine learning approaches, time-series predictions, different recommendation systems, and evaluating prediction techniques used. Preference is given to scientific journal articles and conference papers that present detailed description of the methods used.

It should be noted that pieces that are unrelated to research goals are excluded from the main analysis. Moreover, if the contribution of certain publications is insufficient in terms of explaining the methods used or describing the applications realized without sufficient relation to the content analysed or prediction techniques described in the review, it should not be taken into consideration. Duplicate records must be excluded in case multiple databases indicate presence of the same publication.

### D. Literature Analysis and Synthesis

The literature selected is classified into thematic segments dealing with aspects of traditional machine learning, time series prediction, deep learning, and recommender systems. Every study is reviewed in terms of the goal, suggested method, area of application, evaluation issues, and limitations. For example, studies on predicting online video popularity create background for predicting at content level [1] and studies on streaming recommender systems and methods based on deep learning help to research personalized content discovery [2], [3], [4].

A qualitative analysis is performed in order to determine common patterns in the methodology of conducting research and identify differences in goals and limitations that can somehow hinder the implementation of the aforementioned methods in OTT dataset analysis. The synthesis also takes into account the difference between predicting popularity and predicting shifts in popularity trends.

### E. Scope and Limitations of the Review

This review is mostly concerned with what methods and challenges can help analyse metadata of films and television. It must be stressed that data concerning actual viewing records, subscriber information, and internal data about their users from streaming services is usually out of reach. Thus, it is necessary to proceed with caution with respect to the conclusions about the relevance of the methods.

The structure of this paper is aimed at creating a framework for collecting, processing, analysing, and visualising big data. Therefore, the paper does not make any claims about being a systematic review and does not provide quantitative results.

## III. OVERVIEW OF OTT CONTENT ANALYTICS

OTT content analytics is all about gathering, processing, analyzing, and interpreting movie and television program-related data that is shared through digital streaming services, which would help in analyses regarding the characteristics of certain content and spotting trends, stains, and changes in popularity over time. As it is based on data analysis, it can help in content planning and trend spotting, as well as the construction of future models.

OTT content analytics is as effective as the data accessibility and data quality and relevance. By metadata such as titles, release dates, genre, original language, popularity, and rating, one could gather useful information. The TMDb API provides movies and TV series metadata like release information data, genre identifiers, popularity numbers, as well as rating statistics.

### A. Characteristics of OTT Content Data

OTT content analytics is all about gathering, processing, analysing, and interpreting movie and television program-related data that is shared through digital streaming services, which would help in analyses regarding the characteristics of certain content and spotting trends, stains, and changes in popularity over time. As it is based on data analysis, it



can help in content planning and trend spotting, as well as the construction of future models.

OTT content analytics is as effective as the data accessibility and data quality and relevance. By metadata such as titles, release dates, genre, original language, popularity, and rating, one could gather useful information. The TMDb API provides movies and TV series metadata like release information data, genre identifiers, popularity numbers, as well as rating statistics.

Another very important characteristic is the dynamism of popularity information. According to TMDb, the popularity score is driven by several factors, such as voting activity, views, watchlist activity, learn about releases, and previous scores with differences for each media type. Hence, popularity score should not be a substitute for the number of viewers or a direct measurement of the streaming demand in general.

### B. Role of Exploratory Data Analysis

Exploratory data analysis is used to understand the relations, distribution, and structure of the database, prior to creating predictive models. It can find the duplicates, the missing values, and issues in the data. This method is quite useful in the analysis of OTT content as it can demonstrate the distribution of the films or TV shows according to their year of release and language. The distribution will depend on the type of the content and its genre and popularity effectiveness rating of each piece of content can be shown graphically. Furthermore, EDA can find the correlation among numerical data but do not provide a reason for the correlation result.

EDA is beneficial when determining the new limitation of data. As no content containing low number of votes has more stable average ratings than the content with the high number of votes. In the same way, the fact that some movies were released without particularity mention of their release date influences the accuracy of the output of the analysis.

In the current research, the significance of EDA is the innovative aspect that allows for discovering trends in the analysis, making it possible to give the most valid results.

### C. Data Preprocessing and Quality Management

The process of data preprocessing is crucial in the preparation of OTT metadata for its analysis. The data obtained from different endpoints in the API or CSV files can have several duplicate entries, missing values, inconsistent format for dates, and discrepancies in the name of the attributes. If these problems are handled improperly, errors may ensue in both the statistical analyses and the predictive modelling of the data.

Preprocessing could involve several different steps such as:

1. **Data integration:** Merging all TV and movie information into a master data file.
2. **Duplicate handling:** Looking for duplicates through the combination of different identifiers and keeping useful records.
3. **Missing-value treatment:** Assessment of missing values and adoption of suitable strategies like keeping missing values, imputation when necessary, or discarding unsuitable data points.
4. **Data-type standardisation:** Transformation of dates, numerical values, and categorical variables into correct formats.
5. **Validation:** Identification of consistent identifiers, validation of the date format, numeric range validation, and validation of the content-type field.
6. **Documentation:** Recording decisions made during preprocessing for improved replicability and further analysis.

The missing values cannot be filled in with zeroes as zeroes can have numeric meaning. Similarly, duplicates cannot be identified by title alone, as different works might have the same title. After the preprocessing stage, the master data set needs to be validated to make sure that the records and attributes are right for further analysis and engineering. These steps help create a solid data basis for the suggested OTT content analysis framework.



### D. Feature Engineering for Content Analysis

Feature engineering involves modification of already available features in order to make the representation more suitable for statistical modeling and machine learning. The choice of the features has to be guided by the goal of the prediction, as well as the availability of the data and the potential for introduction of unavailable at the time of prediction information.

In terms of OTT metadata, the temporal features can involve release year, month, and quarter. The categorical features can be content type, original language, and genre. The numerical features can involve popularity, vote average, and vote count. Textual data can be transformed into numeric features as well, if the research involves text-based modeling.

The genre features can be turned into multi-hot features in order to correctly identify titles with several genres. The categorical features can be encoded with an appropriate method, while numerical features can be scaled.

Feature Engineering should also tackle the problem of data leakage. If the goal is to predict the popularity of the movie at the time of its release, then the data collected after the movie's release should not be considered as a feature unless otherwise stated in the problem statement. Subsequent votes or other measures of popularity may have information that is not available at the time of prediction. Hence, a well-defined definition of the target to be predicted and the prediction horizon needs to be done before selecting features.

Finally, the set of features is to be determined through exploration and reasoning about the domain. Different sets of features can then be tested for their effectiveness in predicting.

### E. Applications and Limitations of OTT Analytics

The use cases for OTT content analytics could include content popularity analysis, examining release-patterns, exploring trends across genres, identifying language-specific comparisons, and building predictive models. Such analyses could help to identify properties of available con-tent and investigate what metadata characteristics are associated with popularity.

Popularity prediction and time-series forecasting tasks are related but have some differences. The aim of popularity prediction is to estimate the expected popularity of individual titles based on their available attributes, while the time-series forecasting task predicts a value of a target variable from observations ordered in time. A forecasting task, therefore, requires an appropriate target variable that varies over time and sufficient historical observations.

Another application of OTT content analytics is recommendation systems. Unlike popularity prediction at a content level, recommendation systems aim to find relevant items for specific users or contexts [2], [3], [4]. A metadata-only dataset could be used for content-based analysis or similarity-based recommendations, but for personalised recommendations, additional information about users, interactions, or preferences would be necessary.

Despite the described applications, it is critical to note several limitations associated with the use of the discussed types of public metadata. First and foremost, the data set does not include information on the duration of the viewing of videos, their completion, subscriptions, or other measures specific to a given platform. Second, popularity scores and ratings should also be considered with caution, as their collection methodology may affect their accuracy, and therefore these scores should not be used as a direct indicator of audience interest. Finally, it is essential to recognize that the dynamics of individual preferences, methods of distribution, and the platform's catalog may vary significantly over time.

Accordingly, the framework presented above will be used to analyses the available movie and television metadata, and specific predictive methods will be developed and tested on the appropriate data sets to identify the correct statistical indicators. Ultimately, it is essential to ensure that the conclusions do not go beyond the capabilities of the data sets used and the goals of the proposed predictions.



## IV. MACHINE LEARNING APPROACHES FOR OTT CONTENT PREDICTION

Machine learning offers various techniques for analysing content metadata, estimating popularity measures, identifying patterns, and building predictive models to facilitate data-driven forecasting. The decision on what method to apply depends on the type of predicted target, the properties of the input data, the number of observations, and the chosen validation strategy.

In particular, there are two major aspects of OTT analytics that rely on predictive models. The first is related to content-level popularity prediction, which refers to the estimation of a specific target value for a movie or TV show based on the available features. Trend forecasting is the second direction, which implies the prediction of a trend of a certain indicator over time, e.g., the average popularity level of content within a particular month or year. Both types of predictions involve different sets of input variables, target variables, and validation techniques.

The literature review on video popularity prediction contains numerous studies considering the application of hybrid machine learning models for the estimation of the popularity of a particular product upon release [1]. The research on recommender systems and deep learning in video platforms enables a better understanding of how users find videos of interest [2], [3], [4]. The next section will discuss the main directions of popularity prediction and trend forecasting in greater detail.

### A. Traditional Machine Learning Methods

Traditional machine learning algorithms are frequently applied to prediction and classification tasks involving structured sets of data. They can be suitable for OTT metadata as many of the available attributes including release dates, genre categories, original languages, popularity scores, voting stats can be presented in numerical or categorical feature forms.

**Linear Regression:** Linear regression estimates continuous targets by finding the relationship between the target and one or more features. It is a good starting point to predict some numerical feature related to popularity but might struggle to capture more complex interactions.

**Decision Trees:** Decision trees partition the data according to feature criteria. They can capture non-linear dependencies and have relatively good interpretability. Their main issues are instability and overfitting potential.

**Random Forest:** The Random Forest algorithm integrates the outputs from multiple decision trees. It identifies patterns that are not linear and shows how features influence one another - those capabilities make the algorithm appropriate for organized OTT metadata - but the success of the model relies on how researchers choose features, adjust parameters and confirm results.

**Gradient Boosting:** The Gradient Boosting method creates a series of models one after another to minimize errors in forecasting. It is efficient for data that exists in tables plus for complicated connections between the characteristics of digital media. The hybrid method to predict popularity researched by Jeon et al. offers a specific framework for this type of media analysis [1].

Traditional machine learning models can provide a good baseline to compare more complex mod-els against. The best algorithm to use will need to be found experimentally.

### B. Time-Series Forecasting Methods

Time series forecasting uses data that has been ordered chronologically to predict future values. This can be used within OTT analytics if there is an appropriate metric to track at regular intervals. You can use this to predict how much content you have each month, total releases per year, average popularity per release period or what percentage of your content falls into certain categories.

**Moving Average:** Smooths out short run fluctuations giving a general indication for the series. Can be used to give a basic forecasting method but often reacts to abrupt changes slowly.

**Autoregressive Integrated Moving Average (ARIMA):** Uses previous observations and differs the data. This can be used if the dataset has been taken consistently over an adequate period of time and the assumptions are reasonably appropriate.



**Seasonal ARIMA (SARIMA):** An extension of ARIMA to include seasonality in the model. This could be a useful method to use if there are seasonal trends within the chosen indicator and there is enough history of the indicator.

**Exponential Smoothing:** This provides different weighting to previous observations. The method can include level, trend and if the right variation is used seasonality. This could be a good method to use to forecast aggregate OTT indicators over time.

The choice of time series method will depend on how the variable behaves over time. If the dataset is only one observation per movie/television show then it isn't automatically a time series. However, it could be aggregated to create an appropriate time series or combined with other appropriate time-based information.

To ensure accurate evaluation the training data should come before the validation and test data. If the time-based data is randomly distributed between all subsets, then future information may enter the model and cause overly positive estimates.

### C. Deep Learning Approaches

Deep learning is an approach to learning representations from a dataset using multi-layer neural networks. It can be beneficial if you have a large enough dataset, and your prediction task requires complex feature interactions.

**Feedforward Neural Networks:** These networks aim to find non-linear relationships between the input features and the target variable. They could be used to predict how popular certain content would be based on their structured meta-data, by comparing them to simpler baselines.

**Recurrent Neural Networks (RNNs):** These networks aim to process sequential observations. They can therefore identify temporal dependencies. They could be used for time series forecasting if there is enough historical data.

**Long Short-Term Memory Networks (LSTMs):** These are a form of RNN designed to find long term relationships within sequences. They could be used for forecasting temporal OTT indicators, but they aren't guaranteed to outperform simpler models if the time series is short.

**Deep Learning Based Recommender Systems:** The idea behind these recommendation methods is to learn representations of users, items and interactions. There has been much research into deep learning-based recommender systems, and Zhou et al. provide a survey of the main categories of methods [3]. For personalisation of OTT recommendations, they will require some interaction or preference data.

Whilst deep learning can capture complex relationships, it can require more resources and fine-tuning. Just because a model is complicated does not guarantee it will perform well. Therefore, any deep learning models should be compared against simpler models with the same target, train/test split and evaluation metrics.

### D. Recommender Systems

The overall aim of recommender systems is to find items which are likely to be relevant to a user or certain context. In particular, for streaming services, users can find new movies or television content from large content libraries. Unlike popularity prediction, recommendations aim to find items which are relevant to users. Highly popular items may not be the most relevant to all users.

**Content-Based Filtering:** Content based filtering finds similar items based on item attributes. The content attributes could include genres, language and descriptions in the OTT dataset. This method can be used with just the metadata of the items however this would not give any personalisation based on the user's actual interaction.

**Collaborative Filtering:** Collaborative filtering finds items which are likely to be interesting based on the interactions between users and items (e.g. ratings/watching). This will depend on how much interaction data there is available. Metadata alone will not allow traditional collaborative filtering to be implemented.

**Hybrid Recommendation:** Hybrid recommenders combine two or more methods to overcome limitations of each method. For example, combining content-based features with user interaction data.



There is research on the Netflix recommendation system which discusses the algorithms used for recommendations, and how this adds value for businesses [2]. There is general research on using deep learning and deep reinforcement learning for recommendation tasks [3][4]. However, this will require data and evaluation methods to fit the recommendation goal.

For the proposed OTT project, it would be possible to carry out similarity analysis based on metadata. However, if the user interaction data cannot be found, a fully personalized system could be considered an extension.

### E. Comparative Overview of Approaches

Various machine learning techniques require different resources, have certain advantages and disadvantages. The aim of the following table is to give an overview of different methods to help choosing one. This is not based on any experimental results of the proposed project.

Table I: Comparison of Machine Learning Approaches for OTT Content Analysis

| Approach | Primary objective | Potential input data | Main strength | Key limitation |
|---|---|---|---|---|
| Linear Regression | Continuous-value prediction | Structured metadata | Simple baseline and interpretation | Limited representation of complex nonlinear patterns |
| Decision Tree | Prediction or classification | Structured metadata | Interpretable decision rules | Can overfit |
| Random Forest | Prediction or classification | Structured metadata | Captures nonlinear relationships | Less interpretable than a single tree |
| Gradient Boosting | Prediction or classification | Structured metadata | Models complex feature relationships | Requires parameter tuning |
| ARIMA or SARIMA | Temporal forecasting | Chronologically ordered target values | Provides statistical forecasting baselines | Requires an appropriate time series and model assumptions |
| Exponential Smoothing | Temporal forecasting | Historical time-series observations | Useful for level and trend patterns | May be unsuitable for complex irregular patterns |
| RNN or LSTM | Sequence prediction | Sequential observations | Can learn temporal dependencies | Higher complexity and data requirements |
| Content-Based Filtering | Item similarity and recommendation | Content metadata | Can operate without a full user-item interaction matrix | Limited ability to learn actual user preferences |
| Collaborative Filtering | Personalised recommendation | User-item interactions | Learns from interaction patterns | Requires sufficient interaction data |
| Hybrid Recommendation | Personalised recommendation | Metadata and, where available, interaction data | Combines complementary information | Greater implementation and evaluation complexity |

*Source: Conceptual overview of the reviewed methodologies [1]-[4]. Note that no quantitative*



*information about the performance is provided, neither is there any general hierarchy.*

Overall, classical machine learning techniques offer practical baselines for structured OTT metadata. Time series techniques allow prediction of future changes in aggregated metrics. Deep learning can be an option if the data and prediction task justify the increased complexity. For recommendation approaches content discovery can be achieved, but may require additional information beyond the meta-data.

When choosing the final model for your framework, consider what target variables are available, how far ahead you want to predict, the computational requirements, and results from the validation process. Aim to keep the comparison between models transparent.

## V. REVIEW OF EXISTING RESEARCH

A range of research fields have been undertaken related to content analytics in OTT. These include predicting content popularity, recommender systems for streaming, deep learning-based recommenders and adaptive recommenders. This section will look at four chosen pieces of research and discuss their relevance to our proposed OTT content analysis and prediction framework.

### A. Popularity Prediction for Online Video Content

This paper proposes a hybrid machine learning framework for predicting the popularity of new content on online video streaming platforms. This is achieved using a combination of XGBoost and an embedding based neural network [1]. The paper shows the potential using multiple learning techniques to predict the popularity of content at an individual level.

This research can be applied to OTT analytics. It shows how Machine Learning can be used to predict individual content popularity. It gives a foundation to understand how content metadata and predictive variables can be used to help predict content popularity.

The results should be considered within the scope of the data used to build the model. If the model was built on specific viewing data from the service itself, this will not necessarily translate to a dataset of mainly metadata. The presence of real viewing data, how popularity was calculated and when the data was collected can all impact if this method is applicable to other datasets.

This paper provides a good foundation for the proposed framework. This will provide a good insight into how well-structured data can be predicted. If this was to be adapted to TMDb metadata there must be a clear dependent variable and independent testing.

### B. Recommendation Algorithms in Streaming Services

This paper by Gomez-Uribe and Hunt [2] covers the recommendation system used within Netflix. It discusses the algorithms used, the business goals and how the recommendations are assessed. The purpose of recommendation algorithms is explained, along with how off-line experimentation and A/B testing can be used to improve recommendations.

It clearly shows that there are many factors which must be considered when implementing recommendation systems. This includes the algorithm chosen from machine learning, but also the goals of the business, quality of data available and how the algorithm is assessed. Quality could be assessed using user engagement or retention figures.

This is useful as it explains the difference between recommending personalised content, and simply predicting what content will be popular. Whilst a title might have high popularity scores overall, it may not be a good recommendation for each user. A less popular title may be more relevant to an individual user's preferences.

However, the system used by Netflix will use specific data and testing infrastructure to the platform. This means that we won't be able to fully replicate the testing methods without the right data relating to user interactions and a good testing environment. The paper does provide good concepts to help us understand how to define our goals and differentiate between analysis using the metadata and personalisation.



### C. Deep Learning for Recommender Systems

A detailed review of deep learning-based recommender systems was presented by Zhou et al. [3]. This covers how deep learning is becoming more prominent in recommender systems and categorises methods that can be used to learn representations from the data to model complex user-item relationships.

This is relevant to OTT platforms due to the size of content catalogues and the variety of user preferences involved. If there is the right amount of interaction data available, deep learning methods can be used to help with the recommendation tasks. This gives a wider overview of the methodology behind learning-based recommender systems.

However, deep learning is not always the best option for an OTT platform's analytics tasks. It will depend on how much training data is available, what kind of representation you have and the type of metrics being used for evaluation, along with how complex the prediction task is. If only metadata is available, then whilst we can understand similarity between contents, we lack the information to learn personal preferences.

This helps to show why deep learning could be considered within the proposed framework. But if the extra complexity isn't justified by the available dataset and the prediction task defined then it shouldn't be implemented.

### D. Deep Reinforcement Learning in Recommender Systems

[4] Chen et al. provide a review of Deep Reinforcement Learning for Recommenders. An overview of why reinforcement learning could be used to assist with recommendations is presented, alongside categorisations of current literature. New trends and challenges are also discussed.

Reinforcement learning allows an agent to learn how to make sequential decisions based on the feedback received. This can be helpful when looking at recommendations as it enables recommendations to be made which adjust over time.

This paper helps to expand the area of research for recommendations by showing the importance of making sequential decisions. However, in general, reinforcement learning requires an appropriate environment for the agent to interact with, as well as feedback and a clear reward objective. These are not provided with just a set of movie and TV metadata.

Therefore, using Deep Reinforcement Learning is more of a potential research area for the future. Extra data would be required along with an appropriate environment to evaluate if adaptive recommendations would provide any useful benefits.

### E. Synthesis of Reviewed Studies

The four studies in this review tackle different aspects of this area. The study by Jeon et al. [1], looks at predicting the popularity of new online videos. Gomez-Uribe and Hunt [2] examine how recommendations are determined and evaluated in a streaming environment. Zhou et al. [3] provide an overview of deep learning methods for recommendations and Chen et al [4] look at how Deep Reinforcement Learning can be applied here.

Overall, it demonstrates that predicting the popularity of content and providing recommendations should be considered as two separate tasks. This is because one is predicting how well content will perform whilst the other is determining what would be relevant for users or situations. However, deep learning and reinforcement learning can be used to model this problem. The choice of approach will depend on what information is available and what you are looking to achieve.

A limitation to these areas is that the methods were developed to achieve specific aims, using specific data types. This means they cannot be directly compared. They will differ in how they define their target, what features they use, what metrics they use to evaluate and how they validate. Results gathered from private viewing/interaction data may not translate well when applied to public metadata.

The OTT framework proposes a solution to this by providing a clear process for gathering, preprocessing, exploring, engineering, modelling, evaluating, forecasting and visualising metadata. The methods described by the papers have provided context to how this can be accomplished. However, whether this is effective needs to be investigated experimentally using the data available.



## VI. RESEARCH GAPS AND LIMITATIONS OF EXISTING STUDIES

The literature shows that there are various ways in which Machine Learning can be used for Time Series Forecasting and Recommendation systems. However, in order to implement these techniques into OTT content analysis it is important to ensure suitable data is available, clear objectives for the predictions are set and proper methods for evaluation are implemented. The following are key research gaps that must be considered when implementing the framework.

### A. Inconsistent Definitions of Content Popularity

There are many ways to measure popularity of content. These could include: viewer activity, ratings, voting activity and popularity scores on certain platforms. When creating a framework to predict popularity, it is important to understand what your target variable really means. In Jeon et al. [1] they looked at predicting the popularity of online video content.

If you are using a publicly available metadata set you must make sure your target variable reflects this. For example, whilst you can analyse the TMDb popularity score as a metadata indicator, you would not call it an actual measure of how many people viewed the content. This requires the target variable to be clearly defined in the proposed framework. The predictions made should also be interpreted within the constraints of your target variable.

### B. Data Quality and Missing Metadata

Metadata from public content can often have missing data, duplicates, inconsistent date formatting, incomplete genre data and differences in attributes between movies and television. This can cause problems when describing the dataset or creating predictive models.

Preprocessing is an essential step of data analysis but needs to be done correctly based on the data types. Missing ratings should not be simply set to zero. Duplicates should be assessed based on correct unique identifiers.

This framework provides data integration, validation, identification of missing data, duplicate data management and feature standardisation. The processes should be tested on the dataset in question.

### C. Limited Access to Actual Audience Behavior

Within literature for streaming recommendation systems, it is common to consider users' interactions and platform specific data [2], [3], [4] such as watch histories, ratings, clicks, watch percentage, and user preferences. However, public metadata does not often contain the full dataset of interactions needed to recreate these recommendation systems.

When applying research methods for personalisation to the metadata dataset, only the attribute data can be used. This can help with popularity and similarity within the content but does not fully capture individual viewing preferences.

The aims of this project will therefore focus around analysing the available metadata from movies and TV. Any claims into personalisation or real audience demands would require further data and would need to be validated separately.

### D. Temporal Changes and Data Leakage

Popularity and trends at the category level can evolve over time. This can create problems when predicting future popularity using a model trained on historical data, particularly if there are changes in the underlying content catalogue or audience interests.

Further issues include data leakage. That is, including information that would not be available at the time of prediction in the model training/evaluation. For example, using a popularity score measured after a title has been released as a predictor of initial popularity would create data leakage.

The prediction time, target variable and forecasting horizon must be specified in the framework design. If the prediction requires looking to the future, then chronological validation and suitable feature selection should be performed. The performance of any forecasting should be evaluated on data that wasn't used to train the model.



### E. Limited Comparability Across Algorithms

As varying datasets, target variables, features and metrics will be used in different studies, it will be important to fairly compare the performance of the models based upon these differences. Furthermore, different objectives are addressed in popularity prediction [1] and recommendation system research [2][3][4] so direct comparison is not appropriate.

The proposed project should include a uniform evaluation method. Appropriate baseline methods should be included. All models should address the same prediction problem using similar methods when training and testing. The correct metrics should be chosen based upon the target variable and the objectives of the research.

This will help to identify if a more complex model provides any improvements over less complex models.

### F. Integration of Movie and Television Analysis

The metadata fields will be similar between these two content types however the structure and timing will differ. Typically, movies will have a single release date whilst TV Shows often have several seasons and episodes. They can be referenced using the first air date or another series level variable.

When combining these two content types you want to make sure you have a field specifying the content type. This will ensure that the common variables are validated before combining the dataset.

It may be beneficial to separate out your movies and TV shows if there is a large difference in distribution or timing. Whether to combine or separate your models should be investigated using exploratory analysis.

### G. Summary of Research Gaps

The literature review identifies challenges such as; defining what constitutes a measurable popularity benchmark, dealing with missing meta-data, separating content analysis from personalisation, avoiding temporal information leakage, ensuring consistent model evaluation and ensuring the correct integration of movies and TV data.

The proposed framework can address these challenges by breaking down the workflow into clear sections of data collection, pre-processing, exploratory data analysis, feature engineering, prediction, evaluation, forecasting and visualisation. This will provide an effective way to analyse public OTT meta-data and investigate the feasibility of prediction methods.

This framework doesn't assume that none of these challenges have been overcome in previous research but instead they are converted into considerations which will be overcome during the implementation of this framework.




## VII. PROPOSED FRAMEWORK AND SYSTEM ARCHITECTURE

### A. Framework Overview

This project proposes a framework that can be used to analyse and forecast content trends within Over-the-Top (OTT) platforms. The framework extracts movie and television metadata, identifies patterns, and creates features to enable prediction of content trends. The framework involves collecting the data, pre-processing, exploratory data analysis, feature engineering, creating models, evaluating their performance, and predicting/visualising results.

Metadata from movies and television shows is extracted using the The Movie Database (TMDb) API. This can contain information such as title, release date, genre, original language, popularity, vote average and vote count. Whilst this provides a good starting point to explore content characteristics and relationships between metadata and popularity related attributes, it does not reflect how much audiences have viewed and subscribed to content on certain platforms.



### B. Proposed System Architecture

This design works from start to finish. Movie and TV show data will be collected from the TMDb API, and stored as raw data. We will then clean the data to help with duplicates, blank entries and incorrect formatting. With this we will have a single dataset ready to be used going forward.

An exploratory analysis will be performed to better understand what content we have, how genres are distributed, year of release trends, language used and popularity related characteristics. We will decide which features are relevant and create new features to use within our machine learning model. With an understanding of what we want to forecast and having the correct observations, we will train and test a suitable model. We can then display our analysis using charts and a dashboard.

Figure I: Proposed Framework for OTT Content Analysis and Forecasting

### C. Proposed Workflow Diagram

The proposed workflow consists of the following stages:

**1. Data Collection:** Movies and TV shows data will be collected from the TMDb API based on the desired time period.

**2. Data Storage:** Data will be stored into raw data sets.

**3. Data Cleaning and Preprocessing:** Data will be checked for duplicates, missing values, incorrect date formatting and wrong data types and cleaned according to what is needed for the specific analysis.

**4. Master Dataset:** A clean dataset will be created to hold all movies and TV show data with consistent columns and data types.

**5. Exploratory Data Analysis:** Summary statistics and graphs will be used to understand how the data is distributed, what relationships exist, what genres are present and any trends in the characteristics of the data.

**6. Feature Engineering:** Variables will be chosen, converted or created to give the correct variables for the models.

**7. Model Development:** An appropriate machine learning or forecasting method will be chosen.

**8. Model Evaluation:** Model performance will be assessed using appropriate metrics. The model will be validated with an appropriate method to avoid data leakage.

**9. Forecasting Trends:** The model can be used to predict future values for a clear target variable if enough historical data is available and an appropriate design for forecasting is used.

**10. Visualisations:** Charts, summary statistics or a dashboard will be used to make the analysis and forecast easier to understand.

### D. Dataset Preparation and Feature Engineering

Data preparation will be a key part of this framework as it affects the quality of your data which can affect the reliability of your analysis. This will involve checking that there are no duplicates within the data by checking for duplicate IDs. There will also be checks to ensure there are no missing or invalid data points within the dataset. Data will also be checked to ensure there are no discrepancies between date fields. Categorical fields will also be checked for inconsistencies. Movie and TV Show records will be standardised so they can be used together. A field will be added to indicate the type of content.

The features will be engineered depending on what you are trying to predict. Features could include release year, genre, language, vote average, vote count and historical popularity. New features could include age of content, release time period and genre indicator. All features will need to be checked to ensure they are available and suitable for the model.

Care needs to be taken to avoid any data leakage. For example, if you want to predict something in the past, you wouldn't want to use a feature that was recorded in the future. Also, TMDb popularity scores shouldn't be directly related to audience size. You should make a note on what target variable you have chosen and any limitations.

### E. Model Development and Forecasting

When developing your model, you will need to consider which algorithms are appropriate for your



analysis. In the case of predicting a numerical value using content information you may want to look into creating regression models or tree-based methods. If you want to predict how a metric will behave overtime you may want to look into time series methods. If there is enough data available, the features have an appropriate structure and you have enough computing power, you may want to try neural network-based methods.

Once you've decided on some methods to try, compare them using the same train/test splits and appropriate evaluation methods. In the case of predicting trends in the future, it is useful to have the model trained and tested with sequential splits so that it is being trained on older data and tested how it will perform on new future data. You should choose a model based on its performance.

Also, it is important to note that predicting popularity of content is different to providing personal recommendations. Predicting popularity would be predicting an outcome for an item of content. To provide personal recommendations you would need information about the user's preferences. As metadata does not fully capture how individuals interact with content, this framework will mainly focus on content analysis and predicting trends in popularity.

### F. Evaluation and Visualisation

This stage involves assessing your models to see if they give meaningful predictions when applied to unseen data. For numerical predictions you could look at Mean Absolute Error, Root Mean Squared Error and Coefficient of Determination. As long as the assumptions for them hold. For time series predictions, you should compare against basic baselines.

Visualisation can help explain your findings from the exploration and evaluation. This can be done using things like genre distribution charts, year of release summaries, language comparisons, trends in popularity over time and comparing actual and predicted values. Using a dashboard can help with presenting your findings in an easy-to-understand format.

The reliability of your conclusions will depend on how complete your data is, how well defined your target variable is, how you have evaluated your models and how well they perform. Only make predictions and claims about performance once you have run the relevant experiments.

## VIII. EVALUATION METRICS AND PERFORMANCE ASSESSMENT

### A. Importance of Model Evaluation

Evaluating models is a crucial part of the process of determining if your machine learning model can reliably predict future data. When doing OTT content analytics, you need to take into consideration what type of prediction you are trying to perform, the nature of your data set and whether you have historical observations available. An appropriate approach will help you understand any shortcomings of your model and ensure you are comparing the right models fairly.

When performing model evaluation, it's important to consider whether you are trying to predict a popularity related score for a piece of content, or predict an overall trend over time. This will determine what validation method to use and potentially different metrics. Your evaluation approach should be chosen based on what you want to predict.

### B. Regression Evaluation Metrics

If your model has a continuous numerical target variable, you can use regression-based metrics to measure the differences between actual and predicted values.

Mean Absolute Error (MAE) measures the average differences between the actual and predicted values. The lower this is, the smaller the average error in prediction.

Root Mean Squared Error (RMSE) measures the square root of the average of the squared errors in prediction. Due to the larger errors being weighted more heavily this can show a model that has some large errors.

R-squared (coefficient of determination) measures how well the model explains the variance in the target variable compared to predicting the mean of the target. How this is interpreted will depend on the data you are evaluating. It can be a negative number if the predictions are worse than predicting the mean.

These metrics could be used if you have a numerical target variable, for example if you have a specific



measurement for popularity. However, these should be interpreted in the context of your target variable.

### C. Time-Series Forecasting Evaluation

If you want to predict future OTT content trends, then your evaluation should be in time ordered format. You will use some historical data to train your model, and then leave some out for testing. This will give you a better representation of how your model would perform in a real-world application where you won't have access to future data points.

The performance of your forecast can be measured using MAE and RMSE. If your actuals aren't zero you could also consider using MAPE (Mean Absolute Percentage Error). MAPE can be misleading if there are zero or near zero actuals so should only be used after considering the distribution of your target.

You should also compare your model performance to simple baselines such as the last period, or an average of historical data. Your complex model should only really be considered to be providing any benefit if it improves appropriately over the chosen baseline.

### D. Validation Strategy and Prevention of Data Leakage

A good validation strategy will ensure that you do not have overly optimistic performance metrics. For time-based predictions, it is usually better to split data in time rather than randomly. Rolling origin or expanding window validation can be used for multiple evaluations.

Information from your test set can leak into your training/feature creation. For example, we should not use information about the future popularity to predict the past. Any pre-processing steps which involve learning parameters from your data should be fit to your training data and then applied to your validation/test data.

Your chosen features should reflect what information you will have available when making your predictions. This is particularly important if you are using popularity scores/vote counts etc which change over time.

### E. Comparative Assessment of Models

Where possible, the different algorithms should be evaluated using a common target, time period and validation. Candidate algorithms can include linear regression, decision trees or time series forecasting.

The algorithms should be compared on how accurate they are, how consistent they are across time, how computationally intensive they are, and how interpretable/ appropriate they are to the data. Often, finding the simplest algorithm which achieves an acceptable level of performance is preferable to a complicated model with little improvement.

It is unlikely that one algorithm will give the best results across all OTT prediction problems. The best algorithm choice should be determined based on evidence from your dataset.

### F. Reporting and Interpretation of Results

Your final evaluation report must state what you were trying to predict, what variables you used, what period you trained/tested over, what algorithms you tested, what baseline methods you used, and how you assessed your performance. Results should be presented in an easy-to-understand table with appropriate visualisations such as actual/predicted plots or error plots.

Consider the limitations of your data. For example, TMDb's popularity metrics do not represent actual views, revenue or engagement with subscribers. Unless you have independent data that can provide evidence to show otherwise.

As this project aims to analyse and predict trends in OTT content, your conclusions should be based upon the performance of your model(s). Any numerical results and statements regarding your models being better should only be included once you have run your experiments.

## IX. DISCUSSION, FINDINGS, AND IMPLICATIONS OF THE REVIEW

### A. Discussion of the Reviewed Literature

Machine learning and data analytics can be used to predict content popularity, provide personalisation and identify trends within digital content. There are various methods which may be suitable depending on what you are looking to achieve, what data is



available and what method you will use to evaluate your results.

Hybrid machine learning models can be used to utilise the strengths of multiple algorithms to predict popularity [1]. It has been shown through research into recommender systems that to provide personalisation it is important to consider the users preference, item attributes and interaction data [2] [3] [4]. These models may not be appropriate if you have no access to this information.

Metadata can be used for OTT content analysis to identify trends and patterns within the data. This could include genre, release date, language, vote average and popularity indicators. However, there may not always be a causal link between two variables. You can't reliably infer how an audience would interact with the content based purely on the metadata.

### B. Major Findings of the Review

There are a number of key points to note.

Data cleaning and feature selection is crucial for accurate analysis of your data. Missing data, duplicate data, data formatting issues and changing metadata can impact the results.

Choose the correct type of model based on the type of prediction you wish to make. Regression models may be the correct choice if you are trying to predict numerical target variables, whereas time series prediction models are more suited when trying to predict future values based on time series data.

The task of predicting popularity and making personal recommendations should be treated as separate tasks. Predicting popularity is looking at the content level, whereas personal recommendations require information on individuals.

When evaluating your models ensure that you are using the correct evaluation metric, ensuring that your validation method is consistent and comparing to a baseline. Complex models do not always give you better predictions.

Interpreting trends within your OTT data will depend on the limitations of your data. Using metadata to analyse your data will allow you to see trends within your data, however will not help confirm real viewing numbers, subscriber preferences or platform revenue.

### C. Implications for the Proposed Project

This review provides the basis for the proposed project 'Analysis and Forecasting of OTT Platform Content Using Data Analytics and Predictive Machine Learning'. This project will involve collecting and cleaning data relating to movies and TV shows, conducting exploratory data analysis and finding relevant features in the data.

Overall, the literature has shown that it is important to evaluate the quality of data before building any models. Identifying the aim of the prediction and choosing the correct validation method are key elements when building a predictive model. This can help when choosing the algorithms that can be used.

The analysis could be used to identify historical trends in genres, the release trends, language and how popularity related variables have changed over time. This can aid in making decisions during the exploratory stage of this project and can be used for further research into OTT platform content trends. Claims regarding audience engagement or commercial success would require additional evidence outside of metadata.

### D. Limitations and Future Research Directions

Limitations of this review lie within the scope of the literature chosen and the consistency of the information available. The different data sets, target variables to be predicted, assessment methods and how the results were reported made it difficult to compare them. Some of the studies looked at using recommendation systems made use of user interaction data which may not be available in this project.

Further research could be done looking at alternative data sets, longer time frames for historical data and improved forecasting methods. This could include comparing baseline models with machine learning models to see whether more complex models are actually improving the accuracy. Additional work could look at how the metadata changes over time and what effect that has on the forecasts.

If accurate data is available on views/engagement on the platforms then further studies could be done to validate whether the forecasts made using metadata match up with real audience behaviour. However, until this is possible, the findings should be presented with these limitations in mind.



## X. CONCLUSION AND FUTURE SCOPE

### A. Conclusion

This literature review looks at the role of data analytics and predictive machine learning models in content analysis and prediction within the OTT industry. Several areas were considered including predicting popular content, time series analysis, machine learning algorithms, and recommender systems. Key aspects of successful models such as data quality, feature engineering, suitable model selection and reliable evaluation techniques were highlighted.

The choice of using machine learning depends on what type of predictions you are wanting to make, and what information you have available. Regression can be used to predict numerical content features. Time series techniques can be suitable for analysing trends in content over time. Personalised recommendations is a related area, but will usually require information about the users' preferences/interactions.

The proposed framework provides a systematic approach to gathering and preprocessing OTT metadata, conducting exploratory data analysis, feature engineering, building predictive models, evaluating model performance, and visualising results. This framework needs to be implemented to determine whether it works. Metadata cannot provide information about how people actually watch content or how they perform commercially.

### B. Future Scope

The future of this project is to complete the preprocessing section. Further, identifying a clearly defined prediction goal for this project. Then developing a suitable machine learning and forecasting model. Different candidate algorithms can then be compared against baseline methods using suitable evaluation measures.

Further improvements could include gathering more data, looking at how popularity factors change over time and creating an interactive dashboard to visualise results. Further experimentation could look into how trends differ between movies and TV shows, as well as different genres and time periods of release.

If reliable data on audience viewership or engagement on the platform becomes available, future research could be undertaken to see if our forecasts based on metadata relate to actual audience behaviour. This could increase the real-world applicability of our forecasting of OTT content whilst ensuring clear differentiation between data, predictions and conclusions based on experimentations.

## REFERENCES

[1] H. Jeon, W. Seo, E. Park, and S. Choi, "Hybrid Machine Learning Approach for Popularity Prediction of Newly Released Contents of Online Video Streaming Services," Technological Forecasting and Social Change, vol. 161, Art. no. 120303, 2020. doi: 10.1016/j.techfore.2020.120303.

[2] C. A. Gomez-Uribe and N. Hunt, "The Netflix Recommender System: Algorithms, Business Value, and Innovation," ACM Transactions on Management Information Systems, vol. 6, no. 4, Art. no. 13, 2015. doi: 10.1145/2843948.

[3] H. Zhou, F. Xiong, and H. Chen, "A Comprehensive Survey of Recommender Systems Based on Deep Learning," Applied Sciences, vol. 13, no. 20, Art. no. 11378, 2023. doi: 10.3390/app132011378.

[4] X. Chen, L. Yao, J. McAuley, G. Zhou, and X. Wang, "Deep Reinforcement Learning in Recommender Systems: A Survey and New Perspectives," Knowledge-Based Systems, vol. 264, Art. no. 110335, 2023. doi: 10.1016/j.knosys.2023.110335.