IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

TV Show Popularity Analysis

Rashmi Singh, Srushti Nemade, Atharv Pillai, Binoy Vijaykumar, and Prof. Gayatri Hegde

Department of Computer Engineering, PCE, Navi Mumbai, India - 410206

Abstract*— The television industry is a constantly evolving multi-billion dollar industry. With online streaming services such as Netflix and Amazon Prime, people have access to thousands of TV shows. The rating and reviews that the audience provides is the biggest indication of whether the show is successful or not. With such data available, we can find out what features the most successful shows have in common and the shows of which genre are likely to be more successful with the help of various Machine Learning techniques such as classification and clustering. Algorithms such as k-NN, SVM, Naive Bayes, Decision Trees and Gradient Descent can be employed to build a model with high accuracy. With the worded reviews provided by the audience, we can also perform sentiment analysis using natural language processing to find out what the audience thinks about any particular show. Based on the predictions made by the model we can also make favorable recommendations to different demographics based on their interests.

Keywords— TV show popularity, Sentiment Analysis, Machine Learning

1. Introduction

The number of TV viewers who interact has grown considerably in the last few years – viewers are no longer individual elements. The Web has socially empowered the viewers in many new different ways, for example, viewers can now rate TV programs, comment on them, and suggest TV shows to friends through websites. Some innovations have been exploring these new activities of viewers but we are still far from realizing the full potential. For instance, social interactions on the Web, such as comments and ratings in online forums, create valuable feedback about the targeted TV shows.

2. Literature Survey

A. In the paper titled "Behind the TV Shows: Top-Rated Series Characterization and Audience Rating Prediction", published by Yushu Chai, Yiwen Xu and Zihui Liu used regression models to predict viewer's ratings of TV series based on the existing IMDb database. In particular, the classification model with ratings divided into three subgroups provides the best outcome and is recommended for prediction, although the outcome falls in a relatively wide range. Nevertheless, linear regression using selected features, either by using backward search or PCA, provides improved results compared to linear regression with all available features. The very basic modelused is multiple linear regression. To improve the regression model they also fitted a locally weighted linear model and a reduced linear model by backward search and principal component analysis. However, the error rate was quite high, > 0.3 [2]

B. Tejaswi Kadam et al proposed a system for sentiment analysis in the paper 'TV Show Popularity Prediction using Sentiment Analysis in Social Network'. The process involves data cleaning, data preprocessing, tokenization (the process of infringement a flow of text into words, symbols, phrases, or other meaningful elements called tokens), normalization (eliminating the punctuation, converting the entire text into lowercase or uppercase, converting numbers into words, expanding abbreviations, canonicalization of text, removing stop words from input text data) and Natural Language Processing (NLP). The main drawback though is that it only considers user's comments from social media. [3]

C. In the paper titled 'TV Show Popularity Analysis using Social Media, Data Mining', a predictive model to predict the popularity of TV shows based on user comments from social media is presented by Saura Sambit Acharya et al (2019) [4]. The datasets used to train the ML models were

obtained from IMDB. They used algorithms like "Decision tree", "Random Forest" ,"K-nearest neighbors algorithm" , "Support vector clustering" , "Naïve Bayes classifier" ,"Stochastic Gradient Descent". F1 score, precision score and accuracy was checked for all algorithms by using it on test sets. Out of all these whichever algorithm gave the highest overall score was to be used to predict the statement whether it's a positive or negative. Stochastic Gradient Descent gave the highest accuracy in predictions. [4]

D. "A data mining approach to analysis and prediction of movie ratings" a paper published by M. Saraee, S. White & J. Eccleston of the University of Salford, used IMDb database of around 390,000 movies, television series and video games, which contains information such as title, genre, box-office taking, cast credits and user's ratings. They have found that it is difficult to apply data mining techniques to the data in the IMDb. The data needs extensive cleaning and integration, and this consumed a large proportion of the time available for this analysis. In addition, much of the data is in textual rather than numerical format, making mining more difficult. [1]

2.1 Summary of Related Work

The summary of methods used in literature is given in Table 1.

Table 1 Summary of literature survey

Paper	Advantages and Disadvantages
Yushu Chai et al [2] 'Behind the TV Shows: Top-Rated Series Characterization and Audience Rating Prediction (2015)	Advantages: Very straightforward and simple approach. Disadvantages: Uses a very basic regression model which gives a high error rate.
Tejaswi Kadam et al [3] 'TV Show Popularity Prediction using Sentiment Analysis in Social Network'(Nov 2017)	Advantages: Takes user's sentiment into consideration Disadvantages: Very primitive sentiment analysis framework
Saura Sambit Acharya et al [4] 'TV Show Popularity Analysis using Social Media, Data Mining'(May 2019)	Advantages: Considers user's sentiment, considerably good accuracy. Disadvantages: Scope is limited to comments extracted from viewer's comments from social media websites.
M. Saraee, S. White & J. Eccleston [1], 'A data mining approach to analysis and prediction of movie ratings'(2004)	Advantages: Identified trends in success of movies. Disadvantage: Uses RDBMS approach, simplistic, not time efficient

3. Proposed Work

The disadvantage of the existing system architecture is that it explores only ratings and metadata but does not analyze what users have to say about particular media programs. Here, we argue that text comments are excellent indicators of user satisfaction. Sentiment analysis algorithms offer an analysis of the users' preferences in which the comments may not be associated with an explicit rating. Thus, this analysis will also have an impact on the popularity of a given media show. Thus, the recommendation algorithm will consider both explicit ratings and the output of sentiment analysis algorithms to compute new recommendations.

3.1 System Architecture

The system architecture is given in Figure 1. Each block is described in this Section.

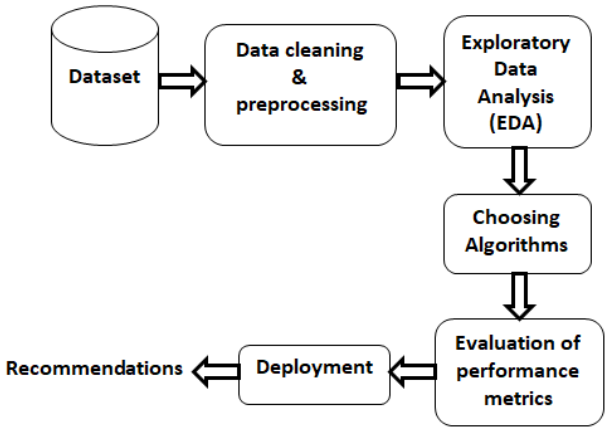

Fig. 1 Proposed system architecture

A. Dataset: A dataset is a collection of data, usually presented in tabular form. Each column represents a particular variable. Each row corresponds to a given member of the dataset in question. We have used the IMDb dataset for this project as it is very vast.

B. Data cleaning and preprocessing: Data cleaning and preprocessing is a process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset to make the dataset suitable for machine learning methods

C. EDA: Exploratory Data Analysis (EDA) refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypotheses and to check assumptions with the help of statistical summaries and graphical representations.

D. Algorithm: It is important to perform EDA before determining which algorithm to choose. The technique to be applied depends heavily on the type of the dataset we have. It is also very advisable to apply many algorithms and then choose the one which provides the best results based on the performance metrics.

E. Evaluation metrics: Performance metrics like F1-score, accuracy, error, log loss, precision, AUC (area undercurve) play an important role in determining the best model and optimizing the model.

E. Output: After the evaluation of the Performance matrix we create a model in which these algorithms try to recommend items that are similar to those that a user liked in the past, or is examining in the present. It does not rely on a user sign-in mechanism to generate this often temporary profile. In particular, various candidate items are compared with items previously rated by the user and the best-matching items are recommended.

3 Requirement Analysis

The implementation detail is given in this section.

3.1 Hardware

Processor	2 GHz Intel
HDD	500 GB
RAM	4 GB

Table 3.1 Hardware details

3.2 Software

Operating System	Windows 10
Programming Language	Python
IDE	Jupyter Notebook

Table 3.2 Software details

3.3 Dataset

IMDb dataset is easily available and it contains a very extensive database of movies and TV shows and is updated daily.

3.4 Evaluation Metrics

The quality of a domain system can be evaluated by comparing recommendations to a test set of known user ratings. These systems are typically measured using precision and recall.

Precision: Precision is the ratio between the True Positives and all the Positives.

$$P = \frac{TP}{TP + FP}$$

Recall: The recall is the measure of our model correctly identifying True Positives.

$$R = \frac{TP}{TP + FN}$$

ACKNOWLEDGMENT

It is our privilege to express our sincerest regards to our supervisor Prof. Gayatri Hegde for the valuable inputs, able guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration of this work. We deeply express our sincere thanks to our Head of the Department Dr. Sharvari Govilkar and our Principal Dr. Sandeep M. Joshi for encouraging and allowing us to present this work.

REFERENCES

1. *M. Saraee, S. White & J. Eccleston*, 'A data mining approach to analysis and prediction of movie ratings' University of Salford, England

2. *Yushu Chai ,Y. Xu, Zihui Liu et al* 'Behind the TV Shows: Top-Rated Series Characterization and Audience Rating Prediction', 2015

3. *Tejaswi Kadam, Gaurav Saraf, Vikas Dewadkar, P.J Chate* 'TV Show Popularity Prediction using Sentiment Analysis in Social Network', Nov. 2017

4. *Saura Sambit Acharya, Ashvin Gupta, Prabhu Shankar K.C* 'TV Show Popularity Analysis using Social Media, Data Mining',May. 2019