# TV Show Popularity Analysis using Social Media, Data Mining

**Saura Sambit Acharya, Ashvin Gupta, Prabu Shankar K.C.**

***Abstract: Television group of onlookers rating is a vital pointer as to prevalence of projects and it is likewise a factor to impact the income of communicate stations through promotions. Albeit higher evaluations for a given program are gainful for the two supporters and promoters, little is thought about the components that make programs increasingly alluring to watchers. So as to think about the prevalence of performers, we consider the quantity of hits gotten by the tweets identified with them on Twitter. In this project we are using three different data mining techniques namely – Decision Tree, Naïve Bayes, and XGBoost. We are comparing each data model with other techniques so that we get the most accurate results. The overall objective of our work is to predict more accurately , which tv show will gain more popularity in the future. Here, we have the option to develop a Graphical User Interface(GUI) that may assist any naïve user in evaluating a show and predict it's success.***

***Index terms: Decision tree, Naïve Bayes, XGBoost, Data Mining, GUI***

## I. INTRODUCTION

There are so many different types of genres for our tv shows. Some of them being action, animation, biography, crime, comedy, drama, fantasy, horror, reality, mystery, thriller, etc which have been agreed upon by us socially over a period of time. Different people like to watch different genres and always have a preference. A show's popularity also depends on what genre it belongs to. For example a thriller might be popular among the younger crowd while a drama will fail for that same crowd. Some recent Indian tv shows that have been aired recently and became very popular are "Breathe", "Ghoul", "Mirzapur", "Sacred Games", "The Final Call". These shows were aired for viewing on several streaming platforms like Netflix, Amazon Prime Video and ZEE5. Sacred Games was viewed by an international audience and received some amazing response. It was critically acclaimed and everyone is already looking forward for it's next season to air. Quite a number of people who watched these shows. Analysis of the popularity of these TV shows will help both the audience and the corporate advertisers. Audience will be able to make choices based on the suggestions about what show to watch next and the corporate advertisers will be able to sell ads based on the popularity of these shows.

We will be working on creating an algorithm which will decide whether a particular show will remain popular in the future or not, based on the comments we receive from Twitter or other blogging sites. We also have a tendency to develop a GUI that will assist the user and advertisers to evaluate a particular show. The success of the model will be predicted by using machine learning models.

## II. RELATED WORKS

### 2.1 Anticipating model of TV gathering of people rating dependent on Facebook

The development of Facebook and other social media websites encouraged more people to share their thoughts. Undertakings saw this and began their very own fan pages for the clients to connect with them and express their steadfastness.

Yu-Hsuan Chang, Chen-Ming Wu, Tsun Ku and Gwo-Dong Chen taken a shot at an anticipating model(2013) that utilized the substance created in the television program fan pages by watchers and Artificial Neural Network to perform conjectures on program evaluations. The exploration developed a program evaluations estimate module dependent on Back-engendering Network. It at that point utilizes trained Artificial Neural Network to play out an appraisals conjecture for up and coming projects.

### 2.2 Motion picture Success Prediction utilizing Data Mining

In 2017, Javaria Ahmad, Amr Yousef and Bill Buckles used data mining processes to extract trends and patterns which can be beneficial for movie success prediction.

The systems were connected on a film database and the information experienced cleaning and mix process before the mining methods were utilized. Information mining manages examples of given information and finding patterns. It recognizes the concealed examples and connections among different factors.

Such connections can distinguish succession of occasions, arrangement and grouping.

They gave a valuable model in this investigation which can brought down odds of disappointment and furnished the partners with certainty and an obvious forecast of progress.

### 2.3 Big Social Data Analytics in Football

A football coordinate is an enthusiastic occasion for the football crew fans. They are particularly appended to the club, city and additionally nation. They show social and social connection to clubs. Nicolai H.

**Revised Manuscript Received on May 04, 2019.**

**Saura Sambit Achary,**,Undergraduate (B.Tech), Dept. of CSE, SRM Institute of Science and Technology, Chennai, India

**Ashvin Gupta,** Undergraduate (B.Tech), Dept. of CSE, SRM Institute of Science and Technology, Chennai, India

**Prabu Shankar K.C,** Assistant Professor, Dept. of CSE, SRM Institute of Science and Technology, Chennai, India

Egebjerg, Niklas Hedegaard, Gerda Kuum and Raghava Rao Mukkamala in 2017 used Big Data Analytics to anticipate fan commitment as far as onlookers and television appraisals. It was workable for them to set up a prescient model for the quantity of onlookers and audience members decently depending on just two control factors to be specific match type and match day. The work had impediments including the way that not many matches were played amid the example time frame, no refinement was made between positive/impartial/negative posts and the social information was accessible for a long time.

### 2.4 Big Data Analytics for Program Popularity Prediction in Broadcast TV

In 2017, Chengang Zhu, Guang Cheng and Kun Wang with Big Data to foresee the prevalence of communicate television programs. With the assistance of the forecast model, a communicate television administrator will most likely upgrade the setup of the system ahead of time by conveying enough transmission and capacity assets to disperse well known projects. Auto Regressive and Moving Average (ARMA) models were connected to genuine follows extricated from YouTube and exact forecasts were acquired. They connected a dynamic time warping(DTW) remove based k-medoids calculation to group programs with comparable fame into four transformative patterns, which can catch the innate heterogeneity of program ubiquity. They assembled pattern forecast models utilizing irregular woods relapse and accomplished higher in general prescient exhibitions.

### 2.5 Group of onlookers Ratings Prediction of TV Dramas dependent on the Cast

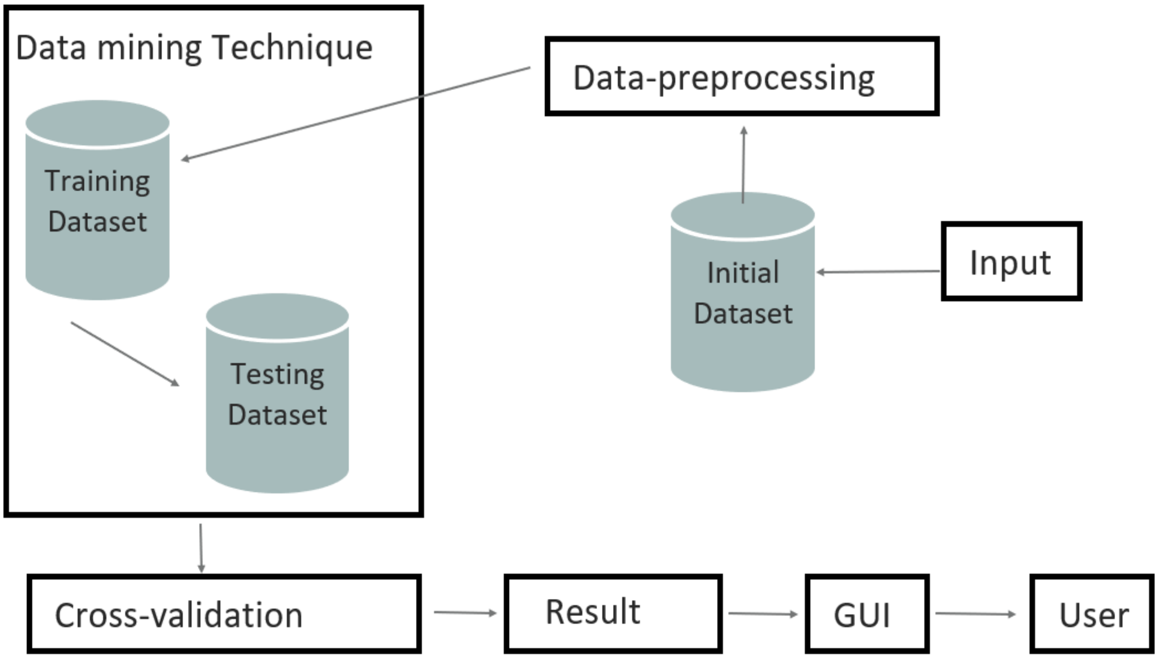
Yusuke Fukushima, Toshihiko Yamasaki and Kiyoharu Aizwa of the University of Tokyo in 2016 taken a shot at an examination to gauge the appraisals of shows before their communicate without utilizing video substance or group of onlookers' reaction to them. They considered the ubiquity of performing artists included, every day perspectives on-screen characters' Wikipedia and notices of them on Twitter were utilized as markers of prevalence. It was clear that a show gets a higher rating if various famous on-screen characters are in it. Anticipate the gathering of people evaluations with high exactness by considering the performers and staff associated with expansion to the traditional data, for example, broadcast appointments and the stations in which they play.

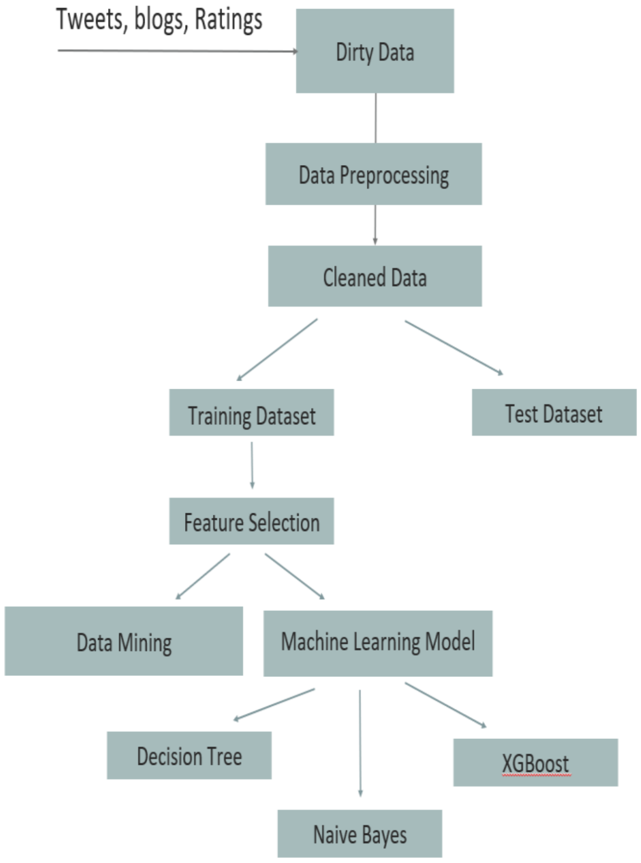
## III. METHODOLOGY

### 3.1 Problem Statement

A predictive model is required to be built which works as a rating system that'll be useful for people who are willing to watch a particular show. They will also be able to get feedback from previous viewers and provide help to new viewers. Vast databases from social media sites will be used for the same.

### 3.2 Architecture





### 3.3 Datasets

Datasets we are using in this project to train machine learning models is Obtained from IMDB site.

This dataset contains motion picture surveys alongside their related parallel opinion extremity marks. This dataset contains 25k review of train sets and 25k review of test sets with their scores. In the whole collection, only upto 30 reviews are taken for same movie since reviews for the same movie will have associated ratings and it contains Entries for reviews with twofold names positive and negative. We incorporate as of now tokenized sack of words (BoW) includes that were utilized in our analysis.

### 3.4 Algorithms Used

Algorithms we used are “Decision tree”, “Random Forest” , “K-nearest neighbors algorithm” , “Support vector clustering” , “Naïve bayes classifier” , “Stochastic Gradient Descent”. We will first check F1 score, precision score and accuracy for all algorithms by using it on test sets. Out of all these whichever algorithm gives us highest overall score we will use that algorithm to predict the statement whether it’s a positive or negative.

3.4.1. Decision tree – A decision tree is graphical structure of all possible outcomes to a decision based on various conditions. It starts with a root node than goes till the leaf node, leaf nodes contains the number of solutions.

3.4.2. Random Forest – This calculation makes a woods with various choice trees. When all is said in done the more trees in the woods the more precise the forecast. It can perform relapse and grouping.

3.4.3. K nearest neighbour—KNN algorithm identifies the k nearest neighbours of any element.it helps us to estimate its class. It can be used for both classification and regression.

3.4.4. Support vector clustering – SVM designs a hyperplane that characterizes all preparation vectors in various classes. Best decision of hyperplane that departs the most extreme edge from the two classes.

3.4.5. Naïve bayes – Naïve bayes algorithm is generally used when we have very less data in our training set. It is used for classification problems, mainly used for text classification involving high dimensional training data sets. For example—spam filtering.

3.4.6. Stochastic Gradient Descent—it is used to build a predictive models. It is used to find optimal solution to a linear regression problem. It involves “loss function”, ”weak learner”, ”additive model”.

F1 score is a formula to compute the score of precision and recall the higher the f1 score is the better prediction will be.

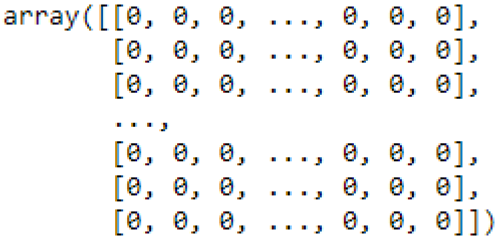
Precision tells us what fraction of your outcome is relevant And recall tells us the fraction of total relevant results correctly predicted by your model.

Stemming is a process in which different forms of word are converted to their root word for ex.
Going, goes ✇go.

Lemmantisation is process in which different forms of words are taken so they can be analysed as a single term by their dictionary form.

Cosine similarity measures the similarity between two elements like in this project it is measuring the similarity between positive words and negative words.

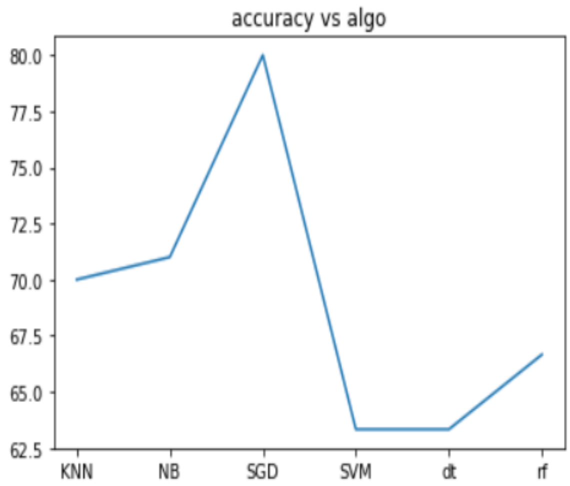
Bag of words is a representation that tells us how many times text comes in different entries. It is used in natural language processing.

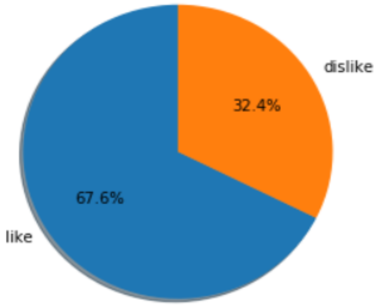


```
array([[0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       ...,
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0]])
```

| | **DT-C** | **RF-C** | **KNN** | **SVM** | **NB** | **SGD** |
|---|---|---|---|---|---|---|
| | BGW | BGW | BGW | BGW | BGW | BGW |
| **F1** | 61.22 | 47.91 | 55.81 | 38.77 | 55.81 | 76.19 |
| **Precision** | 61.11 | 82.75 | 83.92 | 31.66 | 84.92 | 81.98 |
| **Recall** | 61.48 | 54.54 | 59.09 | 50.00 | 56.32 | 74.64 |
| **Accuracy** | 63.33 | 66.66 | 70.00 | 63.33 | 71.00 | 80.0 |

## IV. OBSERVATIONS AND RESULTS





## IV. CONCLUSION

In this paper we have presented a predictive model to predict the popularity of tv shows based on user comments from social media. We are able to obtain significant results over the provided datasets. The model uses sentiments of the viewers and can be implemented for any genre of tv show. Results are highly accurate based on the values we've obtained using data mining and machine learning.

## V. ACKNOWLEDGEMENTS

We thank our entire Department of CSE in our Institute SRMIST and all the people who have previously worked on a related topic. They have indirectly helped and motivated us to develop this model and work on this paper as a part of our Major Project.

## REFERENCES

1. Yu-Hsuan Cheng, Chen-Ming Wu, Tsun Ku, Gwo-Dong Chen. A Predicting Model of TV Audience Rating Basesd on Facebook, 2013.
2. Yusuke Fukushima, Toshihiko Yamasaki, Kiyoharu Aizwa. Audience Ratings Prediction of TV Dramas Based on the Cast and their Popularity, 2016.
3. Nicolai H. Egebjerg, Niklas Hedegaard, Gerda Kuum, Raghava Rao M,Ravi Vatrapu. Big Social Data Analytics in Football: Predicting Spectators and TV Ratings from Facebook Data, 2017.
4. Chengang Zhu, Guang Cheng, Kun Wang. Big Data Analytics for Program Popularity Prediction in Broadcast TV Industries ,2017.
5. Javaria Ahmed, Prakash Duraisamy, Amr Yousef, Bill Buckles. Movie Success Prediction using Data Mining, 2017.