Movie Success Predictor

1st Sameer Thumallapally vthumma@ncsu.edu
2nd Sai Santosh Balusu sbalusu@ncsu.edu
3rd Sai Shruthi Madhuri Kara skara2@ncsu.edu
4th Sruthi Talluri stallur2@ncsu.edu

I. Dataset Description

1) Rotten Tomatoes is one of the most popular film websites, which combines movie information, critic reviews and users reviews. Movies are divided in three categories according to the critics reviews and in two categories according to the users reviews:
   Critics reviews categories:
   Certified fresh: at least 75% of critics reviews are positive and 5 reviews come from top critics
   Fresh: at least 60% of the critics reviews are positive
   Rotten: less than 60% of the critics reviews are positive
   Users reviews categories:
   Upright: at least 60% of the users reviews are positive
   Spilled: less than 60% of the users reviews are positive
   All the records have been scraped as of 07/11/2019.

2) The movie dataset includes 16,638 movies with attributes such as movie description, critic consensus, rating, genre, cast, and all the Rotten Tomatoes scores. The critics reviews dataset includes 930,942 reviews from critics with attributes such as critic publication, critic icon, and review content.

3) Data has been scraped from the publicly available website https://www.rottentomatoes.com.

II. Project Idea

The goal of the project is to predict a movie's success by analyzing existing movies database and data related to the cast and crew associated with the movies in the past. We will be taking both the data regarding the movie such as cast, crew, rating, genre etc. and also reviews associated with the movie given by critics.

An algorithm will be designed to perform the analysis on the reviews attribute of the movies dataset to discretize the content into a binary attribute(positive or negative) to obtain a generalized opinion associated with the movie and is stored as a new attribute in the movies data. The movies data then will be pre processed and we will be performing a KNN classifier to categorize the movies. The movie success to be predicted will be missing the rating attributes, which can not be used, hence we will be using regression model to predict an approximate value.

The idea of the project is to determine how well the model performs in case a predicted rating value for the movie is used instead of the true value, and also determine how addition of review attribute to the movies helps in improving the accuracy.

III. Softwares and tools required

The project will be implemented in python, we will thus be requiring Python 3.0, along with libraries numpy, Scikit-learn, Pandas. We will be using Kaggle in order to obtain the data set.

IV. Work Division

The project focuses heavily on the data preprocessing, due to analysis of the review hence every team member will be involved in this stage. Later we will be working in teams of two, each on different algorithms as proposed, and calculating the performance parameters required for concluding the success prediction rate of the movies.

V. Midterm Milestone

We will be completing pre-processing and cleaning of the data, as we need to combine the data for movies along with its reviews.The reviews for each movie will be processed and assigned a value such as positive, negative or neutral and them combined with the movie data as an attribute. This step requires us to run a sentiment analysis algorithm on the movie review for each movie and later determine a common factor among all and assign a value. The movie data then will be cleaned by removing all irrelevant attributes, and ready to be processed.

References

[1] Javaria Ahmad, Prakash Duraisamy, Amr Yousef, Bill Buckles. Movie Success Prediction Using Data Mining, 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), PP 1- 4.

[2] K Meenakshi, G Maragatham, Neha Agarwal, Ishitha Ghosh. A Data mining Technique for Analyzing and Predicting the success of Movie,2018 J. Phys.: Conf. Ser. 1000 012100

[3] Rijul Dhir, Anand Raj. Movie Success Prediction using Machine Learning Algorithms and their Comparison,2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC).