Bollywood Movie Success Prediction Model using Twitter Sentiment Analysis and Other Features

Project Proposal

Introduction

- Our model will classify movies as Hit, Flop or more, by analysing the recent tweets on the film and calculating a sentiment score, and further combining them with other relevant features for the movie.
- We would also extend the model further to predict the box office collection for the movies either for the opening weekend or lifetime.
- Considering the recent buzz around the impact of social media on the success of movies, twitter would an ideal place for determining this relationship between public sentiment and movie's success.
- In literature, a good amount of work has been done on predicting the success of movies, be it through tweets analysis, review analysis or creating regression models from the movie metadata. In our project we will try to combine tweets analysis as sentiment scores and metadata as features to be used in the classifier and predictor models. We will also try to use some new features like number of clashes in opening weekend, etc.

Data

- The metadata can be extracted from imdb or other bollywood related website using web scrappers. We already have data for 350 films obtained from kaggle with more being available. Filtering of the data based on the availability of other variables will lower the film count. We hope to settle around 100 films.
- Tweets will be extracted using Tweepy or any other third party scrapper which will be searched using some relevant trends for the films. The tweets number is expected to be in range 500 to 1000 or more if necessary.

Baselines to implement

- We will first create a twitter sentiment analysis model that will return a sentiment score (number of positive, negative and/or neutral reviews).
- There are various pre-trained models to try from. We will apply and look for the one giving best results. One of the relevant papers has used PLSA.
- Modifying the sentiment score based on the reach factor of each user, by making a formula.

Teamwork division

- Saksham will look after data managament, exploratory data Analysis and statistical test on the results.
- Deependra will look after model making, coding and hyperparameter tuning, etc.
- Report making part will be done by both.

What is to be done by midway

- Completion of the data collection, filtering and preprocessing part to get the exact number of movies for our training dataset.
- Completion of the sentiment analysis part and using the sentiment score as a feature for the next procedure.
- As an extra exploration, we will try modifying the sentiment score depending on the reach of the twitter user.
- After the midway, we will use our sentiment score feature combined with other metadata features, and create a classifier model using a particular set of features and prediction model for box office collection using some other features.

Expected Result

- We expect the results to improve with the inclusion of the sentiment scores with the other features, and get a good accuracy for the result.

Related Papers

- Gaikar, D. D., Marakarkandy, B., & Dasgupta, C. (2015). Using Twitter data to predict the performance of Bollywood movies. Industrial Management & Data Systems. (Paper on sentiment analysis for Bollywood films)
- Kanitkar, A. (2018, October). Bollywood movie success prediction using machine learning algorithms. In 2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C) (pp. 1-4). IEEE. (Paper on Metadata analysis for Bollywood films)
- Catherine, R., & Chaudhari, S. (2017). Predicting Movie Success from Tweets. (This one used transfer learning approach for sentiment analysis)