Boston University Computer Science

John Spinelli, Shreya Pandit, Cheuk Hin Lee

The Movie Cruncher

Introduction

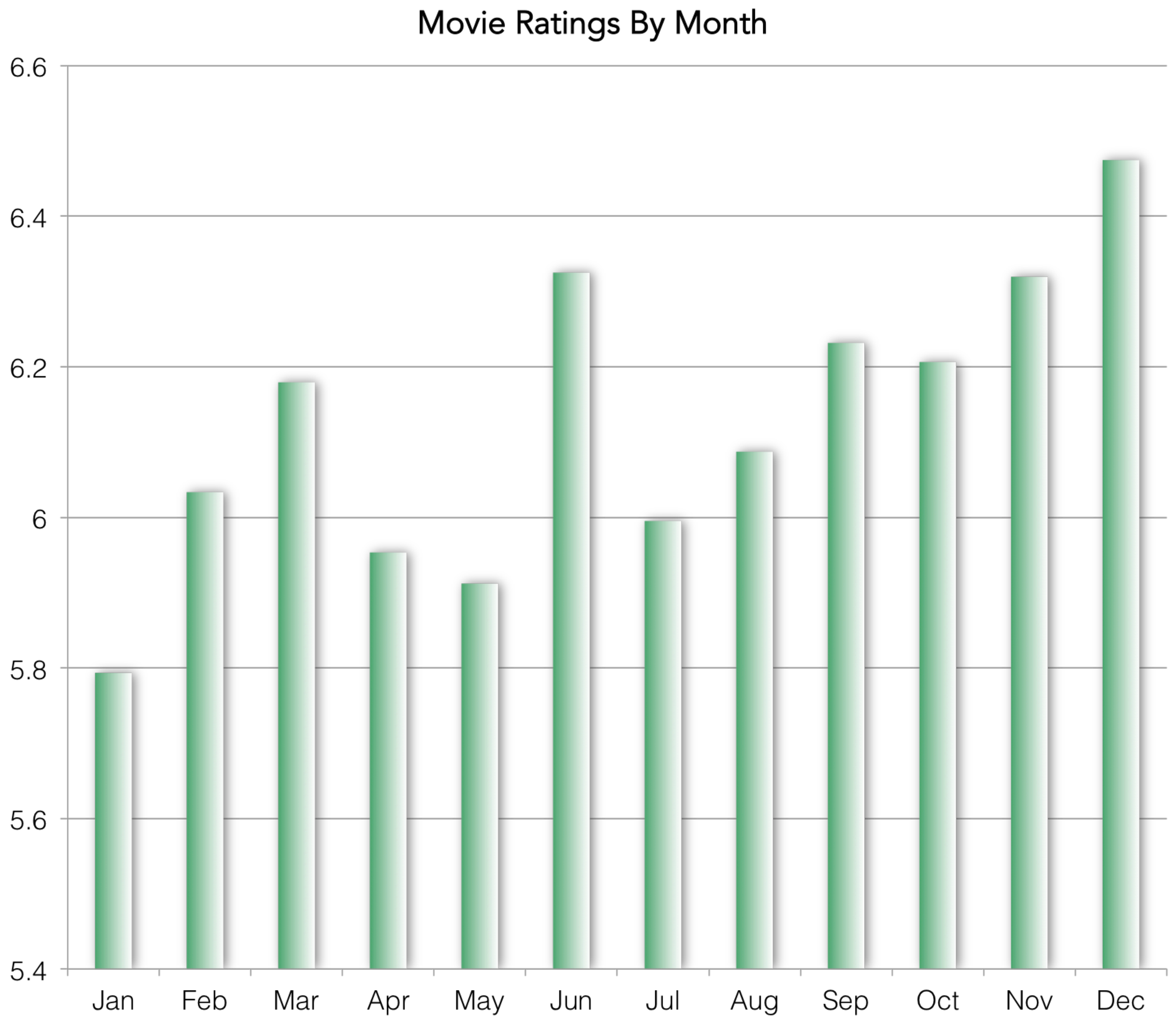
Our project aims to build a prediction engine for movies that helps to identify the important factors that lead to the success of a movie. Using YouTube's and The Movie Database's API and numerous data science techniques, we were able to deduce some interesting factors that helped us determine a hypothesis about movie data and later prove it.

Data Extraction

Using The Movie Database's API, we were able to extract the foloowing fields of movie data from 860 different movies that debuted in 2015 and 2016:

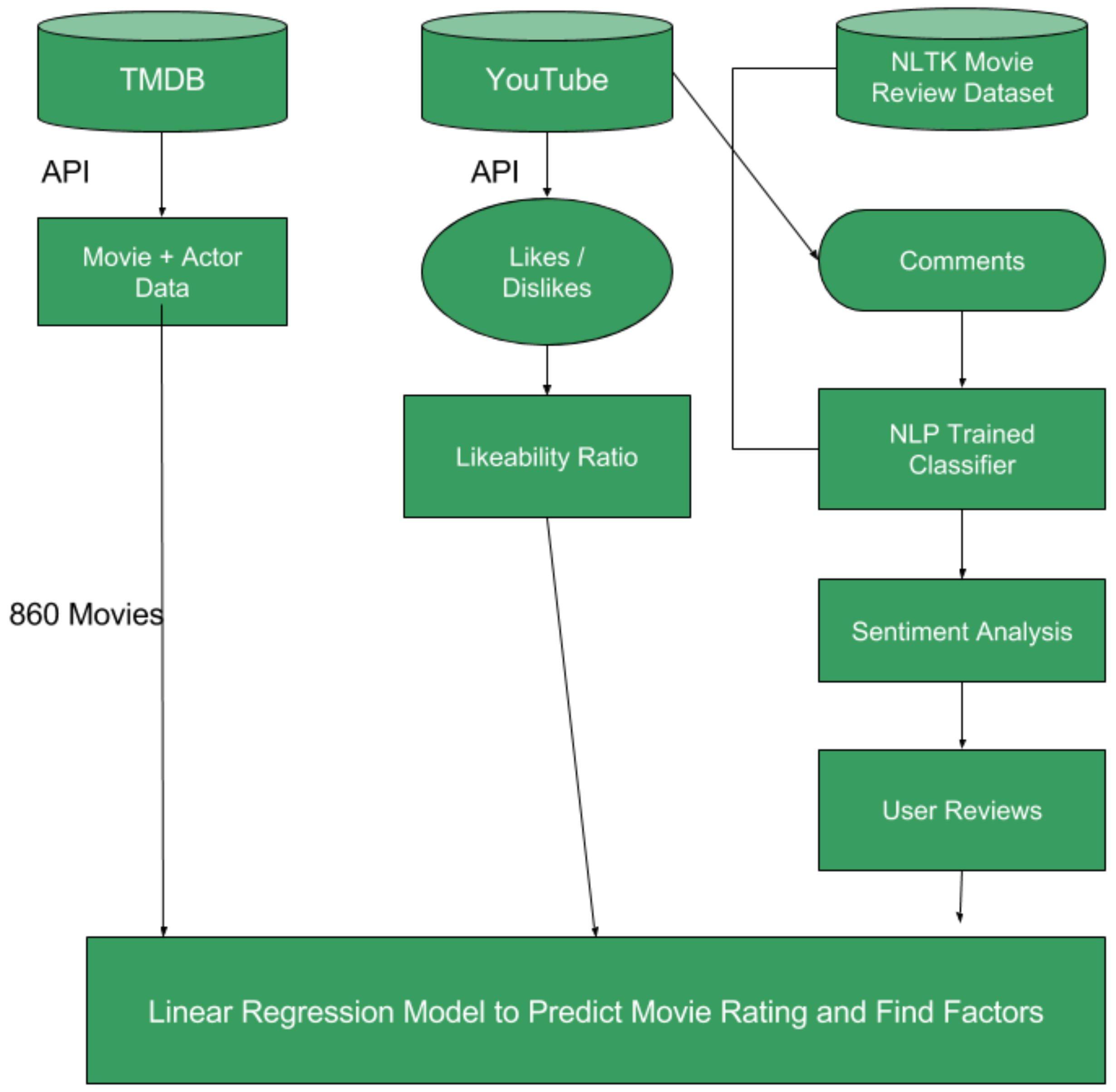
- Popularity, release dates, production countries, textual summaries, revenue, runtime, spoken languages, taglines, vote averages and vote counts.
- Actor data that included actor names and IMDB ID's.

After storing these values in a data frame, we then used YouTube's API to obtain user comments on videos pertaining to the movies we extracted as well as like/dislike counts and view counts of these same videos. With the data obtained from the respective API's, we were able to merge the corresponding features into a data frame that we would use to determine initial trends before modifying any data.



Technique

Sentiment Analysis was done on user comments, which were grouped by movies. It was achieved by feeding data into a classifier trained using NLTK packages. Sentiment Analysis output together with movie data such as genre, time of release and ratio of likes to dislikes was fed into a linear regression model to predict movie rating.


TMDB

API

Movie + Actor Data

860 Movies

YouTube

API

Likes / Dislikes

Likeability Ratio

NLTK Movie Review Dataset

Comments

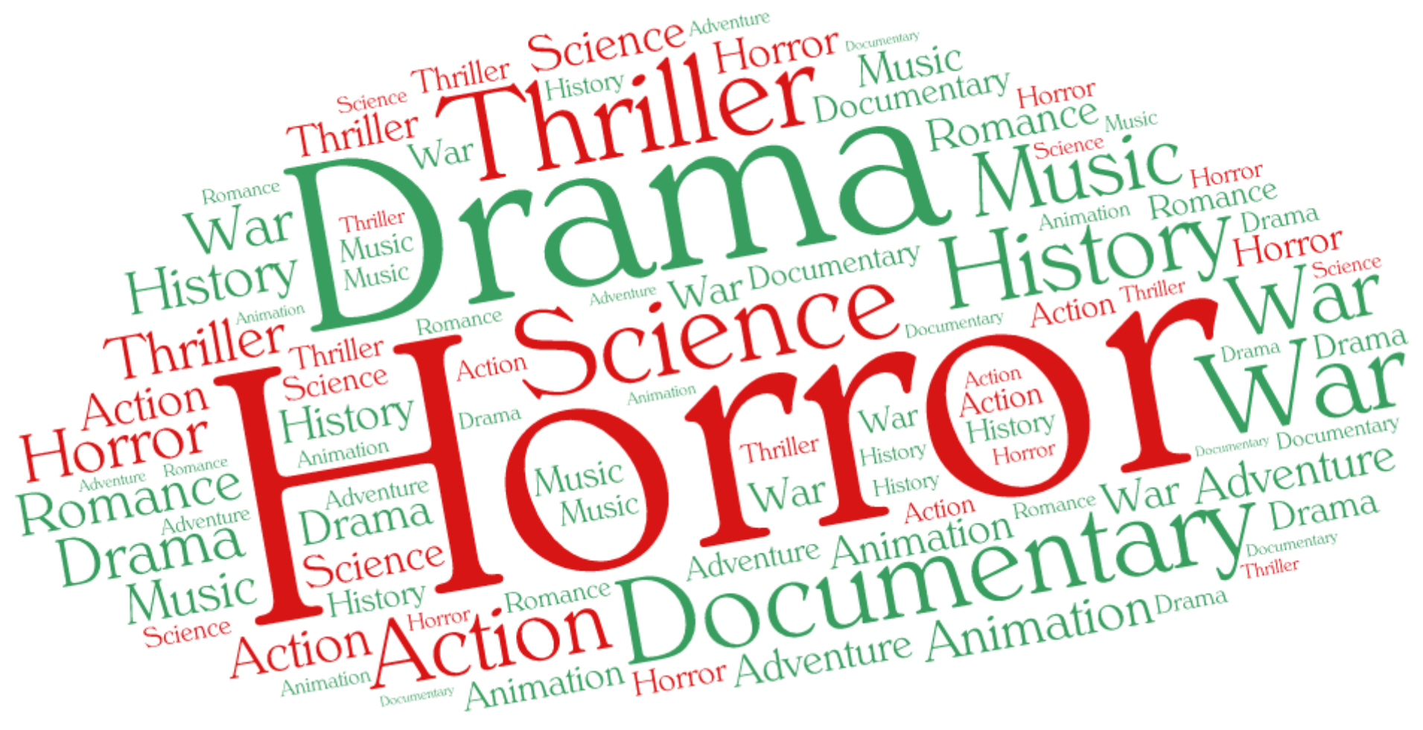
NLP Trained Classifier

Sentiment Analysis

User Reviews

Linear Regression Model to Predict Movie Rating and Find Factors

Metric	Without NLP	With NLP	With NLP+ Likeability Ratio
R-Squared	0.939	0.949	0.950
Adjusted R-Squared	0.937	0.948	0.949
AIC	1604	1446	1428
BIC	1704	1551	1537

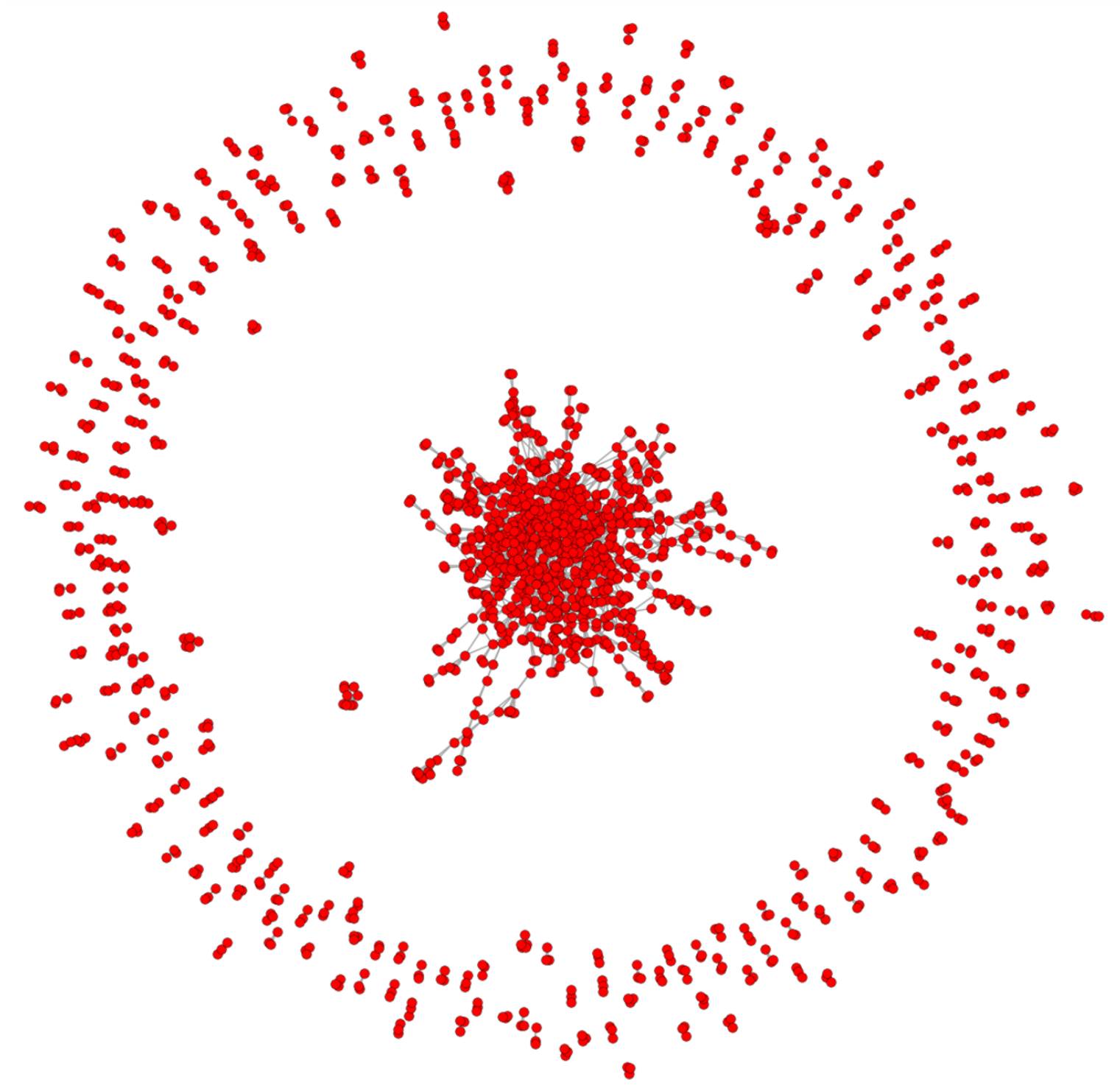


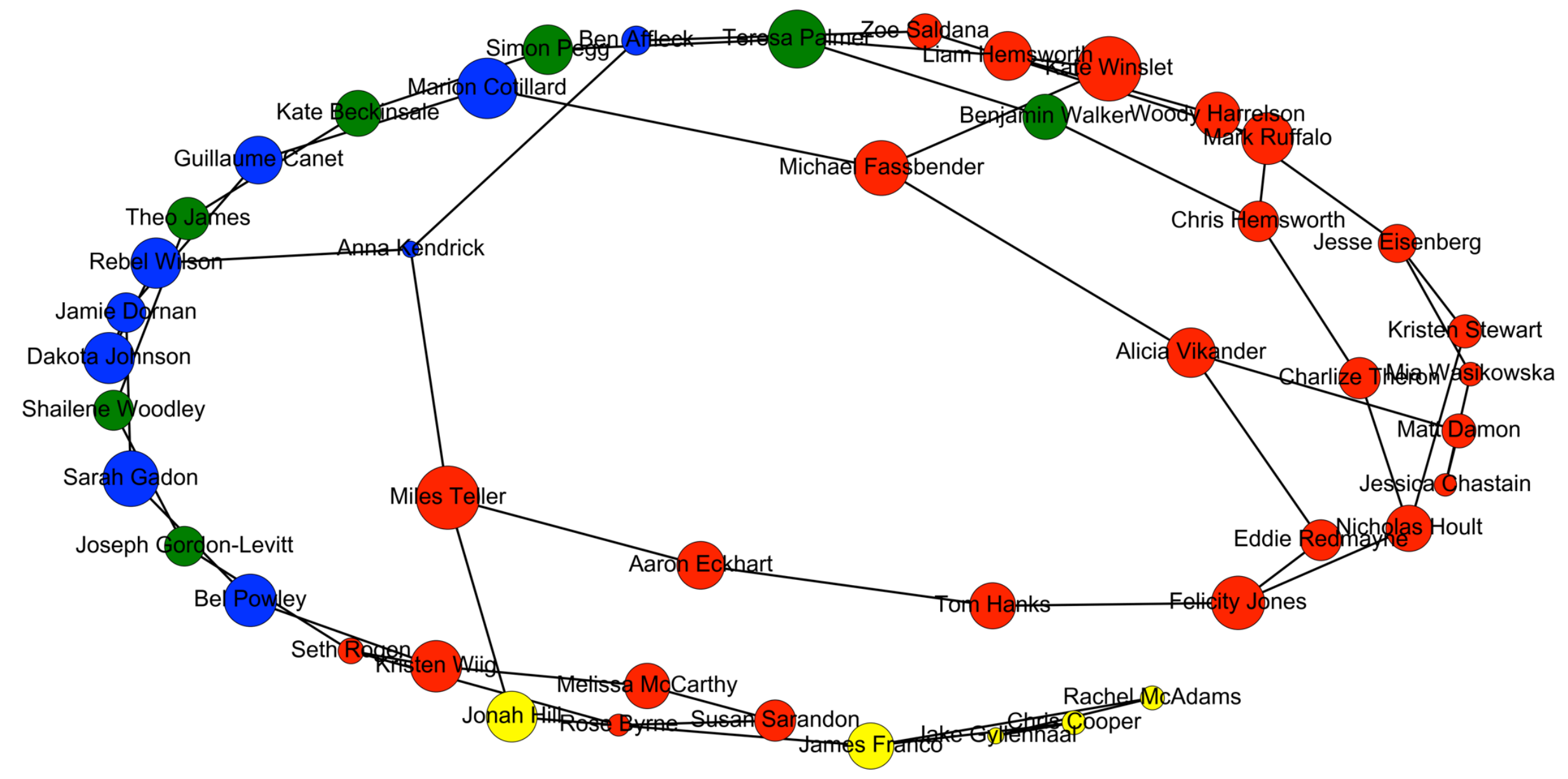
Graph Analysis / Conclusions

Using the Networkx package in Python, we were able to analyze our actor data by performing a K core decomposition and utilizing a Lapachian-based spectral clustering algorithm to identify communities.

Social Media context is important. User sentiment's on a movie's YouTube videos/trailers play an important role on influencing rating. The linear regression model was better fit if sentiment analysis was included in the model.

The results from our linear regression model verify our observations for the trends of movie ratings by release month. The regression also confirms that months like December and June have higher movie ratings while January and April have lower ratings.





Communities -- Label	Mean Vote Average
Red – Actors that mostly participate in action movies	6.4167
Blue – Actors also featured as producers or writers	6.4167
Yellow – Actors that are mostly comedians	6.4121
Green – Actors that at least participate in vampire movies	5.7841