Rating FRIENDS with NLP

Githendra Sagararatne

gsagararatne@gmail.com

Project Problem Statement

The goal of my project is to analyse the FRIENDS tv show script followed by 2 summary datasets to predict key features that can help me predict a successful episode. Natural Language Processing will be used to analyse the script followed by munging the data present in the summary datasets.

Background on Dataset Subject Matter

Based on a bit of kaggle research, I noticed that people only did a bit of EDA on the summary datasets, and built a script generating model with the script dataset. No one combined the 3 datasets and attempted to follow my approach. The reason why I saw this project fit was to view things from a writer's and director's point of view. Identifying certain key elements in a script to predict if a viewer will enjoy the show can not only provide feedback, but also allow the writer and/or director to make changes for the following episodes and/or seasons. From a business standpoint, bring the industry/company more money in.

Details on the source of the data

I was able to get the datasets from Kaggle.

- Summary 1:
    - csv file: 9 columns with 236 rows of data
    - https://www.kaggle.com/rezaghari/friends-series-dataset?select=friends_episodes_v3.csv
- Summary 2:
    - csv file: 10 columns with 229 rows of data
    - https://www.kaggle.com/ruchi798/friends-tv-show-all-seasons-and-episodes-data
- Scripts:
    - txt files: 228 txt files
    - https://www.kaggle.com/blessondensil294/friends-tv-series-screenplay-script

Each row represented an episode, while each column was a feature.

A summary of the preprocessing, feature engineering and any other data cleaning/transformation, and exploratory data analysis (EDA) performed and the motivation and reasoning behind it

I pulled out 14 new features through Feature Engineering. As a result of cleaning data, I was left with 218 episodes out of 236 - a data retention of 92.4%. Even though all the episodes had an IMDb rating greater than 7.2, I tried to identify what features produced a rating greater than 8.4 (the mean). Every rating above said mean was binarised into a *thumbs_up* column.

A summary of all the modelling completed including the process of model evaluation, selection, and results

I initially set up a pipeline and used a cv grid search to identify the best base model from KNN, decision trees, logistic regression and KNN. Logistic Regression was the best base model hence I proceeded to tune its hyperparameters.
The best parameters were: MinMax Scaler, liblinear solver, c=1 and l2 penalty. I attained a 5 fold cross validation accuracy of 52.3%.
Type I error make up for 25% and Type II error makes up for 22.7% of the predictions, Precision and Recall is 54% and 57% respectively.
The CountVectorizer parameters that I set were: stop_words="english", lowercase=True, ngram_range=(1,3) and min_df=0.18.
I was able to replicate the accuracy of the 1426 feature model with 53 PCA components as well.

Findings and Conclusions Based on Data Analysis and Modelling

us_viewers_mill was one of the main positive factors that swayed *thumbs_up*. In terms of character lines - Joey is the most positive. I would give the actor more lines in order to get higher ratings. One of the most negative factors was the number of lines spoken by Chandler. According to my model, the less lines he speaks, the higher the rating. In terms of character (in pairs) plot, the on-screen relationship between Monica and Chandler, and Rachel and Ross seemed to bring higher ratings. In terms of side characters - Janice and Gunther swayed the ratings to be higher. More screen time for them.

A final summary of the business applications and future directions

- Future Directions:
    - Try to salvage the dropped episodes of FRIENDS so I have more data to work with.
    - Try running more machine learning algorithms to get a higher accuracy.
    - Alter the CountVectorizer parameter **min_df** in order to achieve better results.
    - Educate myself with `Rating/Share` and implement it into the mode.
    - Branch out and see things from the viewer's point of view. Get a better understanding of what they like and dislike. Might need to run NLP of reviews to get a better idea.
- Business Applications:

- Directors Kevin Bright, David Schwimmer and Gary Halvorson have promising results.
- Having more than 1 writer per episode produces better results