Arnav Sharma

Movie Review Sentiment Analysis

July 30, 2022

Overview

One subclass of Machine Learning is Sentiment Analysis, which is the process of using computational tools to identify tones expressed in text. As Machine Learning has developed in recent years, there is an active discussion on all the possible ways we can use it. One question which will be explored in this report is: Can Machine Learning be used to rate movies?

Goals

1. **Model:** Create a Machine Learning model that is able to take a comment about a movie and output whether the comment is criticizing or praising the movie.

2. **Movie Ratings:** Collect a handful of comments for a movie and use the model to measure the extent in which the movie is getting praised/criticized. Repeat the process for many movies.

3. **Accuracy:** Compare the model’s ratings for every movie with the Rotten Tomatoes and IMDB ratings for those same movies to see how effective the model is at sentiment analysis.

Specifications

The model will be made using data from the IMDB review dataset, which is a collection of reviews and their sentiment. Only reviews short in length will be used, as reddit comments tend to not be long. Since this limits the dataset to only ~4,500 data points, a Logistic Regression model will be used instead of deep learning.

The movie comments data will be collected from the r/movies subreddit, where for every movie there are comments discussing the movie. There will be 50 total movies taken from the top upvoted posts, so that there are a lot of comments to use for the analysis. In order to avoid very short “meme” comments prevalent in Reddit, each comment must have a minimum of 100 characters. Twenty comments will be collected for every movie, and the mean rating(each rating is either 0 or 1) will be the official movie rating.

In order to measure how good the model is, I will manually collect data of the IMDB and Rotten Tomatoes ratings for every movie that the model rated, and graph all the data in two separate graphs. The graphs will be analyzed to see just how good the model is.

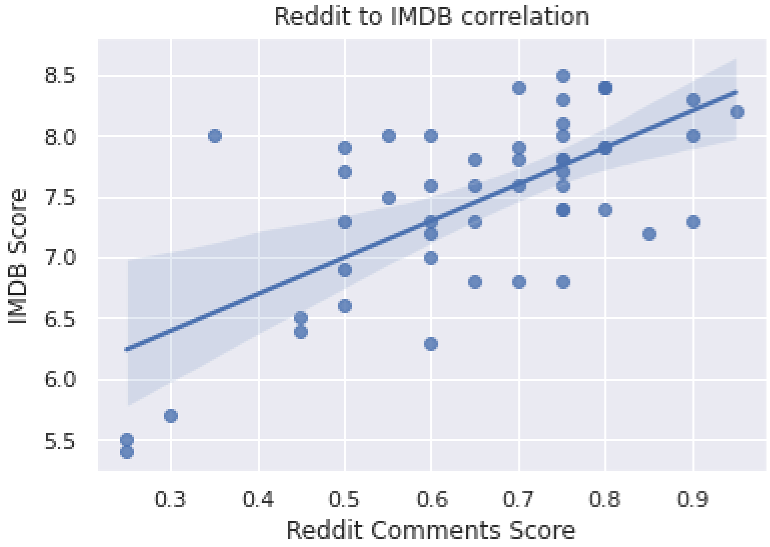
Analysis

1. Model Review Accuracy on Testing IMDB Data

For training, 75% of the data was used, with the remaining 25% of data used for the testing. The model operated by vectorizing the text, then transforming the text into a tf-idf representation, and finally by using Logistic Regression. The model got an 86.4% accuracy on the testing data.

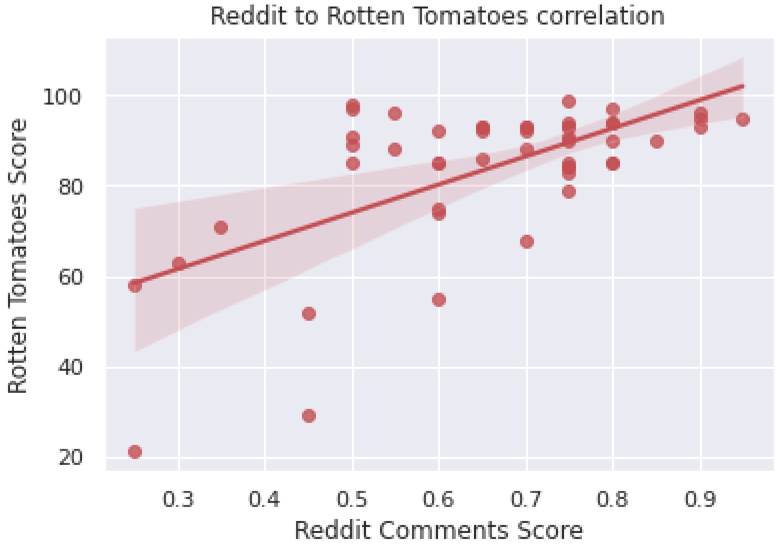
2. Model Movie Rating Accuracy for IMDB

After using the model to rate each movie, a scatter plot was formed using the ratings and IMDB. The model got a correlation coefficient of 0.67.



3. Model Movie Rating Accuracy for Rotten Tomatoes

After using the model to rate each movie, a scatter plot was formed using the ratings and Rotten Tomatoes scores. The model got a correlation coefficient of 0.63.



Conclusion

Overall, Machine Learning shows massive potential for being able to analyze movie comments and rate them, as well as in the realm of Sentiment Analysis in general. Just from ~3,000 data points and without neural networking, the model was able to train itself to rate movies similar to Rotten Tomatoes and IMDB. With more work and data, it is entirely possible to analyze comments of a movie to make a movie rating system better than even IMDB or Rotten Tomatoes.

To be more specific, potential improvements that can be made for this Movie Rating System include:

- More data on movie review comments coming from sites such as Reddit instead of IMDB(which tends to have very long reviews)
- The use of Glove Embeddings so the model can better understand what every word means
- The use of neural network models such as LSTMs instead of Logistic Regression
- Using another model to filter whether comments are proper reviews of a movie or just a joke or unrelated