Jeffrey Wang

Web/Mobile Senior Lab

Period 5 Kosek

13 October 2017

Research Proposal

**Using Movie Ratings to Predict Box Office Success**

**Abstract:**

*A large portion of sales from movie tickets are generated during the first week of a movie’s release. Alongside the rise of social media, sentiment and word of mouth for movies can spread quickly on the Internet. With this in mind, the project will obtain user ratings and box office data from datasets found on the Internet to determine how well can ratings predict the financial success of movies. The project will apply machine learning algorithms to classify movies as a success, neutral, or flop. Classification results for various movies will be provided to users in the form of a web application.*

**Problem/Purpose/Engineering Goals:**

The purpose of this project is to determine to what extent can user-based movie ratings predict the box office performance of movies. A web application will be developed to allow users to view retroactive predictions for various movies released in the past four years. The prediction system will utilize machine learning techniques.

**Background:**

In recent years, machine learning has been rapidly increasing in popularity. Machine learning algorithms can generally be broken into three different categories: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning requires training on a set of variables which allow the machine to predict an outcome variable on data from a testing set.

There are many different classification algorithms that can be used for supervised learning. A few examples include decision trees, support vector machines, and Naive Bayes (Ray).

Machine learning has applications in a wide variety of fields. Some of the most common applications lie under the categories of image recognition and speech recognition. A famous problem in machine learning is character recognition, which trains a machine to learn how to distinguish between letters of the English alphabet and numerical digits. In the task of speech recognition, a program takes in spoken words and represents them as a signal. It looks for patterns in the signal to determine what is being said (S. Sharma).

Additionally, machine learning can be used for statistical analysis to provide predictions for business decisions. One type of analysis that can be performed is sentiment analysis. In the context of business, sentiment analysis can be effective in measuring brand loyalty and predicting consumer trends (G. Sharma). In particular, sentiment analysis can be applied to predict the popularity of movies. A useful measure of this popularity is the total box office gross, which is provided through various services on the Internet.

In recent years, roughly 46% of total sales from movie tickets are generated in the first week of a movie’s release in theaters (Kuizinas). With the rising popularity of social media, word of mouth on platforms such as Twitter can greatly impact the success of movies. A study by the City University of London analyzed micro-blogging word of mouth (MWOM) on Twitter and found that positive messages about movies outnumbered negative ones. However, the negative messages appeared to carry a larger impact in discouraging people from going to the movies than encouragement from positive messages.

Another study from the University of Southern California used sentiment analysis on tweets to predict the box office success of movies. This study was able to classify movies as a hit, average, or flop with 64.4% accuracy (Jain 312). Review aggregators such as Rotten Tomatoes and Metacritic provide Internet users with the opportunity to submit reviews and ratings on movies. As such, user ratings posted online could also be used as a reflection of sentiment towards movies.

**Research Techniques/Methods:**

To begin the project, several movie datasets will be downloaded from the Internet, such as MovieTweetings and MovieLens. These datasets will contain various information about a list of movies, with special focus directed toward ratings, budget, and box office numbers. Python will be the main programming language used to extract necessary information from these datasets. Additionally, there are various modules for Python that can be installed to perform machine learning. From a randomly selected sample of movies, a program will be trained to classify each movie as a success, neutral, or flop.

After data analysis is complete, a web application will be developed to provide the results to the user. The user may select a movie title to get relevant information about that movie such as user ratings and box office numbers. The predicted classifier will also be displayed to the user.

**Materials:**

* MovieTweetings dataset
* Other movie datasets (e.g. TMDB 5000, MovieLens)
* Web Development service (e.g. Cloud 9, Heroku)
* IDLE as the Python development environment
* Python machine learning module (unsure which one)
* Personal laptop

**References:**

Jain, Vasu. "Prediction of Movie Success Using Sentiment Analysis of Tweets." *The International Journal of Soft Computing and Software Engineering*, vol. 3, no. 3, Mar. 2013, pp. 308-13. Accessed 11 Oct. 2017.

Kuizinas, Gajus. "The Twitter Effect." *Go 2 Cinema*, 27 Sept. 2017, go2cinema.com/stories/the-twitter-effect-8. Accessed 11 Oct. 2017.

Ray, Sunil. "Essentials of Machine Learning Algorithms (with Python and R Codes)." *Analytics Vidhya*, 9 Sept. 2017, www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/. Accessed 11 Oct. 2017.

Sharma, Gargi. "Applications of Sentiment Analysis in Business." *ParallelDots*, 24 Dec. 2016, blog.paralleldots.com/technology/deep-learning/applications-of-sentiment-analysis-in-business/. Accessed 11 Oct. 2017.

Sharma, Sheetal. "Top 9 Machine Learning Applications in Real World." *Data Science Central*, 7 Aug. 2017, www.datasciencecentral.com/profiles/blogs/top-9-machine-learning-applications-in-real-world. Accessed 11 Oct. 2017.