CS 579 Movie Project

Ryan Ellison
Richard Hui

I. Introduction

As a moviegoer, you do not have many opinions on how good or bad a movie is, especially on opening day. Up to that point, all you have trailers and predictions on web forums. After that, you can wait for reviews on websites like Rotten Tomatoes or by word of mouth. The main thing to keep in mind is that all those reviews are not always on the quality of the movie itself. There is a mixture of those biased for and against the movie. Unfortunately those reviews are the ones people mainly read. Basically the clash between the reviews and how well a movie does in the box office greatly influence if a movie is worth watching.

The ratings seen on websites like Rotten Tomatoes tend to be biased, and it takes a few days for enough reviews to be posted for the reported rating to stabilize. On Rotten Tomatoes, a good (fresh) score is given to those movie who rate above a certain percentage otherwise a bad (rotten) score is given. It is easily possible for users to purposely rate a movie a certain way to influence the score. Therefore we proposed an experiment to see if we could use the Twitter Streaming API to collect tweets from moviegoers and from there do sentiment analysis to quickly come up with a rating that is close to those seen on movie review websites after their ratings stabilize.

II. Hypothesis

By analyzing sentiment contained within tweets regarding movies playing in theaters, we can provide a quick and accurate movie rating tool that captures a greater amount of feedback than formal, user-contributed review sites such as Rotten Tomatoes.

III. Data

In addition to collecting tweets mentioning “movies” for training of a sentiment analysis model, this experiment consisted of specifically collecting tweets referring to the following movies: *Interstellar, The Hunger Games: Mokingjay, John Wick, Penguins of Madagascar, and Big Hero 6*. In Table 1, ratings from each of these movies are listed from Rotten Tomatoes, IMDb, Metacritic, an average of these three rating sites, and our own “Twitter movie rating” score. Tables 2 and 3 show the top weighted positive and negative terms, as found by our model with logistic regression. Those terms are used to determine the rating score.

	Interstellar	Mokingjay	John Wick	Penguins of Madascar	Big Hero 6
Rotten Tomatoes	73%	66%	84%	68%	89%
IMBD	89%	73%	78%	74%	83%
Metacritic	74%	64%	67%	53%	75%
Average	78.67%	67.67%	76.33%	65%	82.33%
Twitter	86%	83%	75%	90%	90%

Table 1

Positive	
Good	1.08
Love	1.04
Again	0.91
One	0.86
Favorite	0.6
Loved	0.56
Wow	0.53
Day	0.5
Hot	0.5
Wonderful	0.49
Great	0.48
Watch	0.47
Was	0.45
RN	0.43
Cool	0.40
Funniest	0.40
Who	0.40
Both	0.39

Awesome	0.38
Actually	0.37

Table 2

Negative	
Why	-0.90
Some	-0.74
But	-0.72
To	-0.71
They	-0.71
This	-0.70
Didn’t	-0.69
That	-0.68
Don	-0.68
Movies	-0.67
From	-0.64
In	-0.64
Fuck	-0.62
Asshole	-0.56
On	-0.54
Damn	-0.53
Have	-0.51
Waste	-0.51
Not	-0.48
Most	-0.48

Table 3

IV. Methods

First, tweets were gathered that referenced either the word “movie” or a number of current movie titles, making use of the Twitter Streaming API. About 1500 tweets were collected in this fashion for use as training and test data. Then the text within each tweet was manually reviewed and labelled as either “positive” or “negative” (or excluded from the dataset where the tweet’s sentiment was ambiguous or neutral). Upon completion, about 200 labeled tweets remained for the training and testing of this project’s sentiment analysis model.

From this labelled dataset, the sklearn’s vectorizer was used to create feature vectors to train a logistic regression model, again with the use of sklearn. The accuracy of the model was tested with 5-fold cross validation with a satisfying result of 0.7. A few hundred unlabelled tweets were then classified by the model and manually inspected to confirm classifications were generally reasonable.

To produce individual movie ratings, tweets were collected referencing specific movies, and sentiment analysis was performed on the text of the tweets using the model resulting from the previous steps in the project. Each individual tweet was classified as either positive or negative, and the collective percentage of positive tweets was used to create a “Twitter rating” score.

As we were interested in unique impressions from Twitter users on movies they see, we take steps to filter out “viral” tweets about movies by removing retweets. Also, we experimented with the effects of filtering speculative tweets containing phrases such as “I heard that the movie was good…” to see if that improved results by focusing only on first-hand impressions, although such efforts proved to be ineffective and were later removed.

Finally, to determine the quality and usefulness of the program’s rating results, we compare the ratings produced by our tweet analysis to those assigned by other rating sites we ultimately aim to complement/compete with.

V. Results

While the ultimate movie ratings produced by our analysis of tweets obviously differ from ratings of other services, it is difficult to conclude our results need adjustment. Even existing rating sites often have widely varying scores assigned to their movies, so an appropriate bias adjustment is not practical. In fact, it appears tweets are more focused on the entire movie experience (e.g., “I had a nice time with my family seeing Interstellar”), as opposed to a formal movie review. As this is the case, our rating may offer a different kind of useful insight into whether or not a movie is likely to be worth seeing. Also, as useful ratings can be determined after only a couple hundred tweets (although collecting more tweets continues to improve results), so a rating can be determined after only a couple hours of collecting tweets.

Looking into the model, itself, it appears to do a reasonably good job of classifying tweets as positive or negative. Looking at some of the highest-weighted negative terms, some of the “negative” words are commonly used in neutral tweets, so some bias toward negative classification may occur. On the other hand, Twitter ratings are already much more positive overall than other rating sources. Ultimately, improving results may require additional manual classification to build a stronger model.

VI. Related Work

Movie Rating Prediction by Sentiment Analysis: Dylan Boliske

VII. Conclusion

From analyzing the tweets we collected, we could not exactly match the ratings found on Rotten Tomatoes and Metacritic. We saw that majority of the tweets were more personal/emotional than an actual movie review, and tweets focused more on the actual experience than on the plot or character development. In addition, we saw that unless the movie was an absolute flop that the tweets tended to be more positive than review sites. However, Twitter ratings are still useful, as they are both more timely and experience-focused than formal review sites.

VIII. Resources Used

- Twitter API
- IPython
- Github
- Rotten Tomatoes, Metacritic, IMDB