Movie Rating Prediction using Reddit Comments

submitted by:

Bhargav Aditya Ayyagari (baa140030)
Aditya Gopalakrishnan (axg144430)
Bhavya Nataraj (bxn140530)
Shruthi Ramesh Nayak (sxn145130)
Manalee Panda (mxp141330)
Anuprita Rao (arr140430)
Madan Rao (mdr140230)

using
PySPARK, NLTK

PROJECT REPORT

MOTIVATION

Most movie reviews ask people to come up with a number (even though sites like imdb allow you to write a review, in the end it asks you to select a number between 1 and 10). Sometimes, the words(/feelings) are not converted to numbers fairly. So we decided to rate movies based on people's comments about them.

DATA SOURCE

https://www.kaggle.com/c/reddit-comments-may-2015

One of the most popular websites where people talk about everything is Reddit.com. We found a dataset for the month of May 2015. The data is in the form of a csv file with the following fields:

created_utc, **ups**, subreddit_id, link_id, name, score_hidden, author_flair_css_class, author_flair_text, **subreddit**, id, removal_reason, gilded, **downs**, archived, author, **score**, retrieved_on, **body**, distinguished, edited, controversiality, parent_id

Data is in the form of rows, each of which contain the comment(body) that we are analyzing. The score(ups - downs) tells us whether that comment was considered correct or incorrect by other people on the sub.

How does Reddit work?

Reddit is a forum where people talk about anything and everything. Reddit is grouped by “subreddit”. A subreddit is a page for people to talk about similar topics. So a subreddit can be something as general as “movies” or “sports” or to something as specific as “rogerfederer”.

Inside the subreddit, people can start discussions. People reply to this discussion. The comments can be upvoted or downvoted.

DETAILED APPROACH

The project has been divided into 3 modules:

Preprocessing, NLP and Analysis.

The division was done to keep modules separate so that they are easy to understand and also to divide the work among the team.

Preprocessing:

This module takes care of filtering the data and sending only the appropriate rows to the NTLK engine. The comments in the dataset can be as big as a 500 word review of a movie or as small as a one word comment such as “Okay”. So it was necessary to find out which comments made sense to analyze.

First criteria for a comment to qualify was that it should talk about a movie. At times, people on these subreddits go off-topic and talk about random things. As such, we

filtered out irrelevant comments and only talked about ones that were mentioning movies. After doing some rearrangement, we send the data to the NLTK module.

NLTK module:
Each row of data passed to the module contained a movie name as the key and the entire row as the comment. For each such row, we extract the comment and pass it to the NLTK engine. For each movie and its comment, a net score is calculated based on the positive and negative contextual inferences generated by the model on the comments for that movie. We trained a classifier for this purpose and we are using that classifier in the function.

Analysis module:
Once we have the <movie, score> available for each movie in the dataset, we are ready to do the analysis.This is then multiplied with the difference in the (upvotes - downvotes). This multiplication is done because its important to understand the context and validity of the comment. Imagine if a person comments about a loved movie like “Star Wars” and says that “Star Wars is the worst movie ever!”. Based on just this comment, the NLTK engine would give negative rating to the movie. But that is unfair since it is just one person’s opinion. Since Reddit is a community, people can like/dislike a comment and such an incorrect comment will be disliked (downvoted). Thankfully, the dataset has included this score (upvotes - downvotes). So the negative comment will be cancelled by the negative score given by the other users and the legacy of the movie lives on!
Finally, the comment scores for all movies is normalized to a 0 to 10 range just like IMDB.

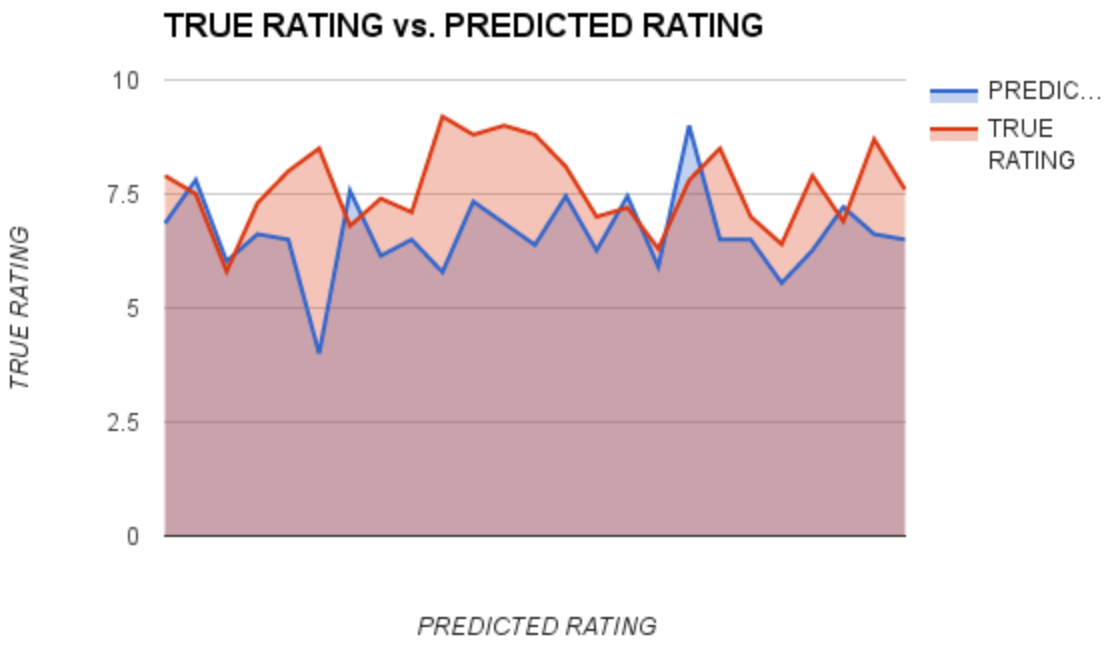
PERFORMANCE EVALUATION

TABULATED RESULTS

MOVIE	PREDICTED RATING	TRUE RATING	YEAR OF RELEASE	DIFFERENCE IN RATING	REASON FOR FAILURE
Iron Man	6.857142857	7.9	2013 (Iron man 3)	-1.042857143	
Harry Potter	7.80952381	7.5		0.3095238095	
Unfriended	6.023809524	5.8		0.2238095238	
Furious 7	6.619047619	7.3		-0.680952381	
Star Trek	6.5	8		-1.5	
Django Unchained	4	8.5	**2012**	**-4.5**	**Only 2 comments and 1 strongly positive but sarcastic and used**

					negative words. Also many times people call the movie as Django, which is hard to catch because that could also refer to other uses of Django
Divergent	7.571428571	6.8		0.7714285714	
The Judge	6.142857143	7.4		-1.257142857	
Transformers	6.5	7.1		-0.6	
Godfather	5.785714286	9.2	**1972**	**-3.414285714**	**More than 5 yrs old**
LOTR	7.333333333	8.8	**2003**	**-1.466666667**	**More than 5 yrs old**
The Dark Knight	6.857142857	9	**2008**	**-2.142857143**	**More than 5 yrs old**
Inception	6.380952381	8.8	**2010**	**-2.419047619**	**More than 5 yrs old**
Avengers	7.452380952	8.1		-0.6476190476	
Mad Max	6.261904762	7		-0.7380952381	
Man of Steel	7.452380952	7.2		0.2523809524	
TMNT	5.904761905	6.3		-0.3952380952	
The Social Network	9	7.8		1.2	
Back to the Future	6.5	8.5		-2	
Thor	6.5	7		-0.5	
A Space Odyssey	5.547619048	6.4		-0.8523809524	
Avatar	6.261904762	7.9		-1.638095238	
The Babadook	7.214285714	6.9		0.3142857143	
Star Wars	6.619047619	8.7	**1977**	**-2.080952381**	**More than 5 yrs old**
Watchmen	6.5	7.6		-1.1	

GRAPHED RESULTS



RESULTS:

The results were accurate within 1 point of the IMDB rating for the 25 movies *(mean: -1.036190476196, standard deviation: 1.2920530107436).*

Most of the error was accounted for because of older movies (more than 5 years old) which did not get a due share of comments because the comments were focussed on current movies and therefore the classics didn't get as much extreme attention.

Also, another exceptional case was *"Django Unchained"* - because the comment used highly polarized, sarcastic, context based language which our engine could not detect and also got a number of upvotes by humans who could understand the sarcasm.

However the results are much more accurate for the other 18 movies apart from these outliers *(mean: -0.43783068782778, standard deviation: 0.7905664155169).*