INFX 547A – Social Media Data Mining and Analysis

FINAL PROJECT – MOVIE MANIA

Nelson Dsouza
Justin Petelka
Shrija Priyanil

INTRODUCTION:

In the age of social media today, it is a boon to marketers, to be able to gauge the reactions of consumers and clients in real time. In this sense, Twitter is probably the most untapped potential source of commercialization for marketing of products, and gleaning enlightening insights. Performing sentiment analysis on tweets from a particular or general demographic could potentially predict sales, consumer reactions and enable real time retrieval of feedback. Using this, we thought it would be interesting to see the relationship between viewer tweets about movies and their ratings on a particular website.

Hence, we decided to ask the following question:
"Does Twitter sentiment for a particular movie (i.e. tweets using the movie's official hashtags) influence or possibly predict a movie's Metacritic rating?"

Accordingly, we created the following null hypothesis:
"Twitter sentiment for a particular movie does not influence a movie's Metacritic rating."

We wanted to disprove this hypothesis – as we researched, we found a paper by people at HP Labs which proved a relationship between social media and a movie's revenue (https://arxiv.org/pdf/1003.5699.pdf). They found that they could efficiently and accurately predict a movie's box office revenue using Twitter, but only after it was released. We extrapolated this finding and assumed that if Twitter sentiment is an indicator of a movie's box office revenues, and could forecast it using sentiment analysis, then a movie's rating must also be correlated with Twitter sentiment.

DATA COLLECTION:

The data was collected using Twitter's REST API and OMDB's API. OMDB, or the Open Movie Database had an easily accessible and free API, which provided ratings from known sources such as IMDB (International Movie Database) and Rotten Tomatoes. Metacritic is a great source for aggregated ratings – aggregated from valid and verified reviews. We retrieved the Metacritic rating using the OMDB API.

We decided to use the Twitter REST API since it's easy to comprehend and use. The drawback with this, however, is that we could only collect tweets for recently released movies. We decided to find the official movie hashtags used for marketing pre and post release for the following movies:

1. Alien: Covenant, released on May 19th, 2017 - #aliencovenant
2. Logan, released on March 3rd, 2017 - #logan

3. Beauty and the Beast, released on March 17th, 2017 - #beautyandthebeast and/or #taleasoldastime
4. Guardians of the Galaxy Vol. 2, released on May 5th, 2017 - #guardiansofthegalaxyvol2 and/or #mantis and/or #babygroot
5. Baywatch, released on May 25th, 2017 - #baywatch and/or #bebaywatch
6. Pirates of the Caribbean: Dead Men Tell No Tales, released on May 26th, 2017 - #POTC5 and/or #piratesofthecaribbean5 and/or #deadmentellnotales
7. Wonder Woman, released on June 2nd, 2017 - #wonderwomanfilm and/or #wonderwoman

The data was collected over a period of a couple of days, and we collected 4000 tweets for each movie. The most recently collected data was for Wonder Woman since we wanted to give a couple of days, post release, for the Metacritic rating to be solidified.

DATA ANALYSIS METHOD AND RESULTS:

Once we collected the tweets for the aforementioned movies, we analyzed the user metadata to ensure the absence of duplicates and potentially spam users. We got a list of unique user IDs and once we were convinced that this was a relatively clean dataset, we moved on to the sentiment analysis portion.

The tweets for each movie were stored in an individual JSON file, from which we first retrieved only the relevant information like the text of the tweet, the user details, favorite count, retweet count and source of the tweet. These were stored in dicts.

We then performed sentiment analysis on the text of the tweet by first splitting the tweet text into words, and only retrieving the words that had a sentiment attached to them e.g. we ignored words like ‘a’, ‘the’, ‘at’ etc. We then compared them to an available list of emotions and words with sentiments to retrieve the relative percentages of all the emotions.

Once we retrieved the emotions present over the collected tweets, we retrieved the Metacritic ratings of the movies.

The analysis for the tweet emotions and the Metacritic ratings were as follows:

	movie	anger	anticipation	disgust	fear	negative	positive	sadness	surprise	trust
0	Beauty and the Beast	3.722084	11.315136	1.538462	4.516129	15.930521	20.918114	4.019851	14.689826	7.940447
1	Baywatch	5.676804	12.104949	2.134764	6.392367	9.230769	28.312463	3.887895	10.995826	6.201550
2	Pirates of the Caribbean 5	4.836538	17.067308	2.278846	4.134615	8.259615	28.259615	3.365385	5.692308	11.846154
3	Logan	3.442825	15.826454	2.459160	4.707536	9.555595	23.766028	3.829264	11.364834	9.836641
4	Guardians of the Galaxy Vol. 2	2.440127	12.245820	9.489381	3.569815	18.481699	21.667420	3.931315	7.614099	8.811568
5	Alien: Covenant	5.273994	10.833069	6.731074	8.758315	12.876148	18.102629	5.685778	5.669940	17.453278
6	Wonder Woman	5.000000	9.805195	2.727273	4.896104	9.753247	28.623377	2.142857	4.025974	18.662338

	Metacritic Rating	movie
0	65	Beauty and the Beast
1	38	Baywatch
2	38	Pirates of the Caribbean
3	77	Logan
4	67	Guardians of the Galaxy Vol. 2
5	65	Alien: Covenant
6	79	Wonder Woman

Now we wanted to find whether there is any correlation between each sentiment and the Metacritic rating. For this analysis, we used the python function “pearsonr” to get the Pearson correlation coefficient and the p-value for each sentiment and the rating. We wanted the p-value to be able to say, with some confidence, whether we would be accepting or rejecting our null hypothesis.

With p-values (lying between the range of 0-1), a value lesser than 0.05 would typically mean that you reject the null hypothesis, and accept the alternate hypothesis, and a value greater than 0.05 would mean we accept the null hypothesis.

Our results for the Pearson coefficient and p-values were as follows, for each sentiment:

EMOTION	CORRELATION	p-VALUE
Positive	-0.39	0.383
Trust	0.43	0.332
Negative	0.32	0.476
Anticipation	-0.38	0.392
Surprise	0.24	0.591
Joy	-0.13	0.78
Disgust	0.23	0.62
Sadness	-0.11	0.805
Fear	-0.08	0.86

As we can clearly see, while the sentiments show a positive or negative correlation to the ratings, the p-values for each are an extremely strong indicator that we must accept our null hypothesis – Twitter sentiment does not influence the Metacritic rating of a movie.

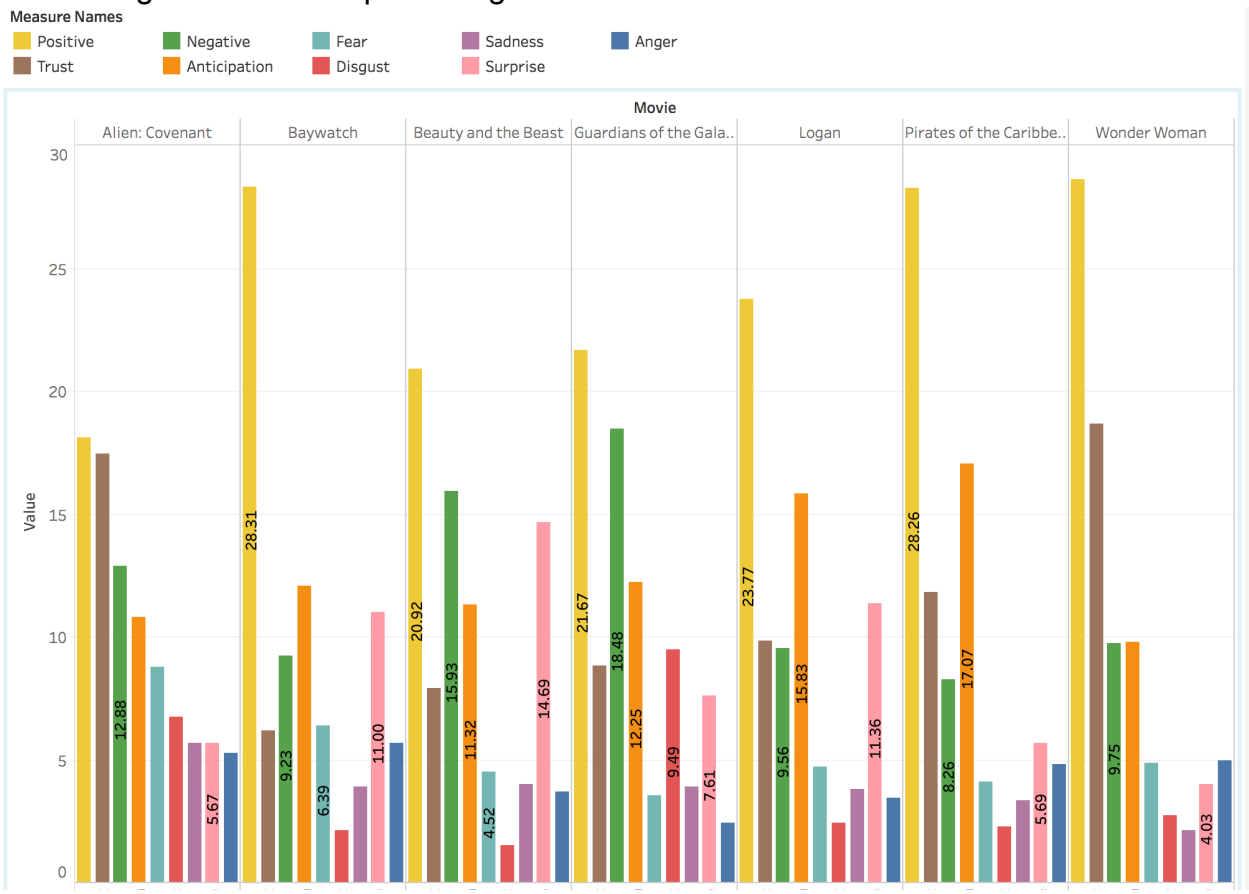
LIMITATIONS:

There were quite a few limitations to our project. The first being that the Twitter REST API only allows for collection of 7 days' worth of historical tweets. The other possible fallacies in our project could be the absence of a large enough sample to conclusively come up with a significant result, the possibility of spam users which corrupted the data and a not extensive enough grammar for our analysis. We also didn't consider bigrams, trigrams or emojis in the tweet texts.

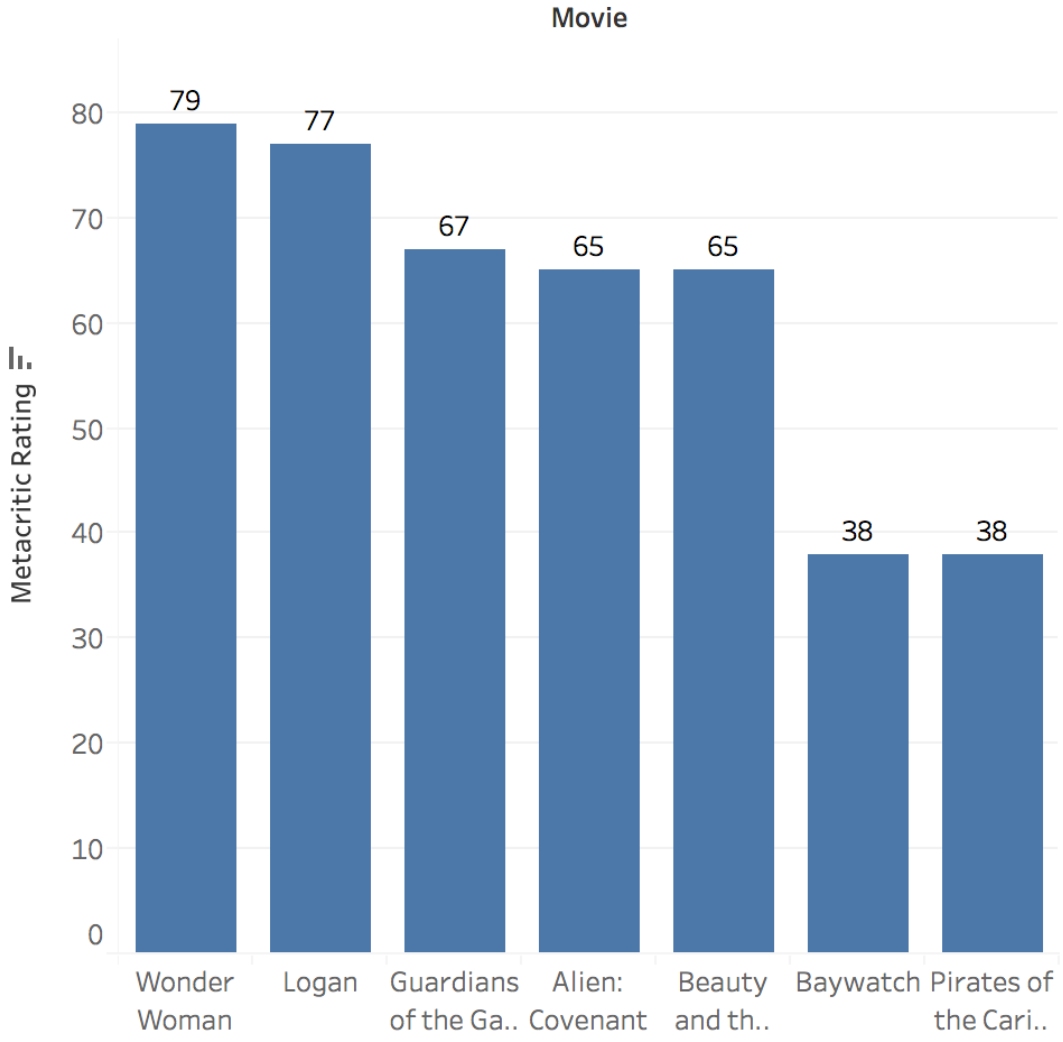
CONCLUSION:

We created some basic visualizations to show our results and its significance:

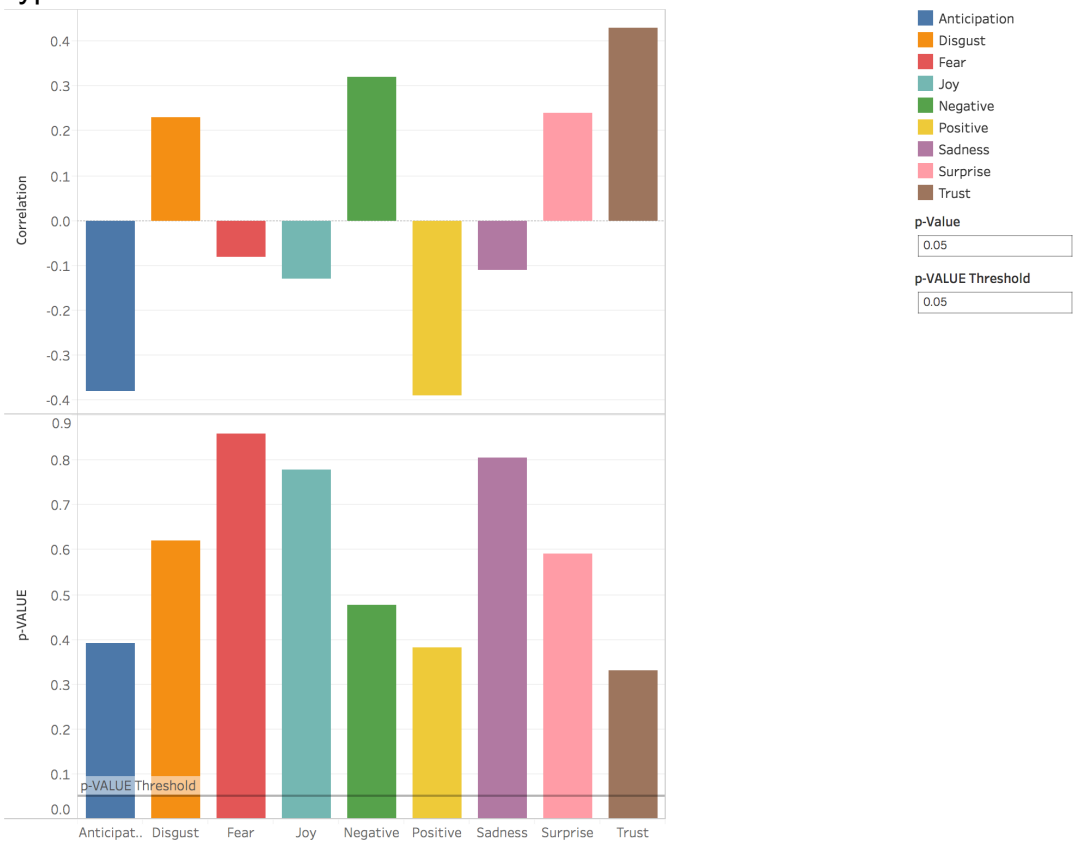
1. Visualizing the sentiment percentages for the movies:



2. Visualizing the ratings:



3. Visualizing the correlation coefficients with p-values for each sentiment's hypothesis:



From these graphs, we can see two very contrasting results:
For Wonder Woman, we can see that the positive sentiment is very high and so is the Metacritic rating, which could lead to the conclusion that there is a possibility of a correlation between Twitter sentiment and the rating. However, if we look at Baywatch, we can see that while the positive sentiment was quite high, the rating is one of the lowest of the 7.

The reasons for this could be many, one being that while a movie could make a high revenue initially due to a big name star cast, director or even both, at the end of the day the story needs to connect to the audience.

Thus, with our limited data, we can say that Twitter sentiment surrounding a movie pre and post release is not related to the Metacritic rating of the movie.