CSCI6250 - BIG DATA ANALYTICS AND SYSTEMS

PROJECT REPORT

SENTIMENT ANALYSIS OF MOVIES UTILIZING LEXICONS AND WEIGHTED ENGAGEMENT SCORES

Daniel Murphy

May 10th, 2020

ABSTRACT

Social media platforms are an ideal source of data for public sentiment analysis. This paper endeavors to leverage the volume and opinion-centric nature of Twitter content to create a methodology for accurately measuring the public sentiment of newly released movies. It proposes that an accurate, unbiased source for calculating the accuracy of movie predictions is the “real-world” data of Rotten Tomatoes Audience Scores, rather than the more common method of using a corpus of tweets with manually assigned sentiment values. By aggregating many thousands of tweets about a movie, it will be shown that weighting the value of each individual tweet relative to the authority of the tweet’s author yields a more accurate estimation of the general public’s opinion of the film.

INTRODUCTION

Discovering public sentiment has long been a subject of interest to researchers across disciplines; from behavioral psychologists studying stigmatizing sentiments across populations to marketing researchers assessing public opinion to inform potential ad campaigns (Shanahan, 2003). Previously, such research was both laborious and problematically limited in scale, requiring teams of researchers to actively search out members of the public, then solicit their opinions with no guarantee of success.

The advent of social media presented a powerful solution to the problems of both labor and scale, with Twitter, the data source for this paper, being one of the largest resources. In 2019, users published over 500 million tweets per day on average and despite being a mature social media platform, usage continues to grow (Andrew Perrin, 2019). The goal of this paper is to find a way to harness the incredible volume of data being output from twitter and use it to derive accurate public sentiment. To this end, it explores various methods for mining tweets and analyzing their sentiment. As a way to focus the scope into a single measurable topic, we concern ourselves with the sentiment analysis of newly released movies.

When analyzing the large amount of textual data generated daily by social media, computer scientists must address two fundamental problems: 1. how to mine it efficiently and 2. how to derive findings from it that are meaningful. Extensive work continues to be done to address the problem of scale; big data eco-systems such as Hadoop have been shown to be quite efficient in mining data generated by the hundreds of the millions of daily tweets (Montoyo, 2012), (Chiplunkar, 2018).

The second problem, deriving meaningful observations from textual social media data, is addressed by Sentiment Analysis. Sentiment analysis is concerned with how an algorithm can successfully assign a sentiment value to textual content. In this context, a sentiment refers to the emotional state or intended opinion of the text's author. Emotions and opinions being inherently subjective and non-binary, their analysis is not something computers are not naturally equipped to address. Fortunately, computer scientists have developed several Natural Language and Machine Learning based approaches to doing just this (Ali Yadollahi, 2017). In this paper, we will take a natural language approach by using lexicons to derive sentiment scores.

Moving forward, this paper has the following structure: Background and Literature Review discusses other work done on analyzing twitter text sentiment. Research and Methodology addresses the tools and methods used in this paper to mine tweets about movies and analyze their sentiment. Results and Discussion covers the experimental findings of this paper and suggests next steps. Finally, Conclusion will conclude this paper.

BACKGROUND AND LITERATURE REVIEW

From the outset, Twitter has been a subject of interest in the field of Sentiment Analysis; and little wonder as the platform provides a reliable volume of data and, by its very conception, is a vehicle for sharing opinions. As far back as 2010, researchers found success using twitter data to conduct sentiment classification using machine learning techniques such as Naïve Bayes Classification (Pak, 2010). In this paper, popular movies were chosen as an ideal subject to measure sentiment, both due to the fact that tweets about movies trend towards subjective opinion rather than observational musings, and because users consistently create a large volume of tweets about movies on a daily basis. These qualities were not lost on previous Sentiment Analysis researchers and there are numerous papers specifically addressing sentiment analysis of movies using twitter data. Amokik, Jivane et al were able to correctly identify sentiment using the vector classifiers: Support vector machine and Naïve Bayes. Their data was trained on tweets that had been manually assigned labels of positive, negative, and neutral (Amolik, 2016). Other efforts to analyze movie tweet sentiment have found success utilizing the K-Nearest Neighbor and Random Forest machine learning techniques; this includes work done by Palak Baid, Apoorva Gupta, Neelam Chaplot (Baid, 2017).

One potential limitation of these machine learning approaches on twitter data is that their accuracy measures rely upon text data that has been manually assigned a sentiment score. One objective of this paper is to find a "real world" measure of public sentiment accuracy that is not reliant on researchers building a corpus of manually scored tweets, and is not based on the twitter platform. Another improvement upon existing studies that this paper pursues is to consider the relative popularity of a tweet’s author when predicting public sentiment. The methods for addressing “real world” measurements and calculating weighted engagement scores based on tweet author popularity are addressed in the next section.

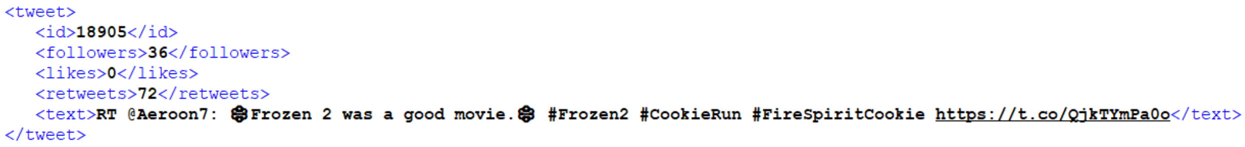
RESEARCH AND METHODOLOGY

Data Collection

Data collection was conducted via the Twitter API using the Java-based Twitter4J library. Using the API's query methods, recent tweets for specified hashtags were searched for. Any data that was returned was then parsed to retrieve the following information:

- User ID of the tweet's author
- Follower count of the tweet's author
- Number of likes the tweet has received
- Number of retweets the tweet has received
- Text of the tweet

Once a tweet's data is downloaded, its text is cleaned to add appropriate escape characters, and finally the data is formatted for output as XML.



```
<tweet>
   <id>18905</id>
   <followers>36</followers>
   <likes>0</likes>
   <retweets>72</retweets>
   <text>RT @Aeroon7: ❁Frozen 2 was a good movie.❁ #Frozen2 #CookieRun #FireSpiritCookie https://t.co/QjkTYmPa0o</text>
</tweet>
```

In total, over 80,000 tweets were mined.

After initial sentiment analysis runs, alarmingly high variance in predictions was encountered. Further investigation revealed that the source of variance was movies having too few tweets to draw on. One movie in particular, *Shooting The Mafia*, yielded only a single tweet, resulting in its predicted score being 73% off. To address these inconsistencies, a minimum threshold of 40 tweets per movie was adopted.

Additionally, term selection turned out to be a significant source of variance. For example, "#Cats" returned tweets that were overwhelmingly found to have positive sentiment scores, when in fact the 2019 film *Cats* was met with disdain by the general public. Looking at the content of #Cats tweets, it was found that the subject matter of many tweets involved heartwarming messages about pet cats instead of opinions about the film *Cats*. Careful attention to hashtags was required to yield useful results- in this case, "#CatsFilm" yielded superior results.

Finally, the titles of some films contain words that are associated with extremely low sentiment scores. Take the case 2019's Best Picture winner, Parasite. Despite being very well received by the public, the presence of the word "Parasite" in tweets about the film resulted in an exceedingly low predicted sentiment score. In the figure below, we see that omitting the word "parasite" from the tweets analyzed results in a much more accurate score.

	Actual Public Score	Predicted Public Score
Tweets with word "parasite"	90	26.725618
Tweets with "parasite" ommited	90	80.242645

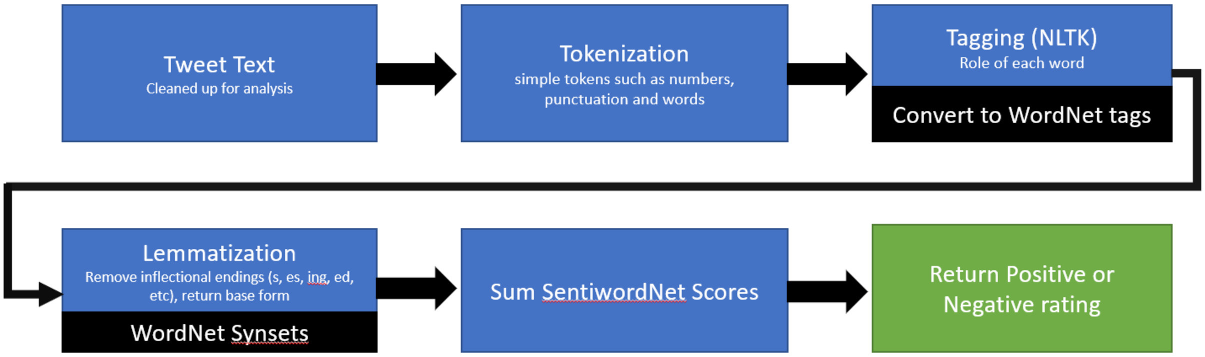
Sentiment Analysis

Sentiment Analysis was conducted on each tweet by using Natural Language Toolkit (NLTK) and the SentiWordNet Lexicon. Natural Language Toolkit is a comprehensive platform for conducting Natural Language Processing in a python environment and was used to convert the xml text data into a Synset format compatible with SentiWordNet. Sentiwordnet is a lexicon based on WordNet, a large lexical database developed at Princeton. In WordNet, words are split into nouns, verbs, adjectives, or adverbs, and are then grouped into sets of cognitive synonyms called synsets. Synsets, in turn, may be interlinked by means of conceptual-semantic and lexical relations. SentiWordNet assigns sentiment scores to each WordNet synset, providing a condensed dictionary that's able to efficiently assign values to over 155,000 English words. A SentiWordNet entry has the following format:

a	00228876	0.625	0	primo	#1	the best of its kind
adjective	ID	Positive score	Neg. score	word	sense number	gloss

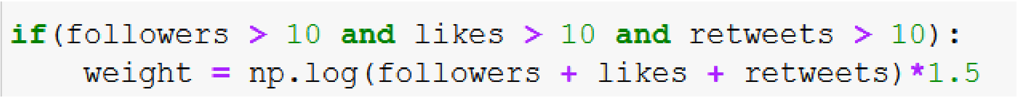
The processing to obtain a sentiment score for a tweet's text is as follows:

1. Text is imported from xml
2. Text is into tokens (words, numbers, punctuation marks, etc)
3. Words are converted into NLTK tags (classifying the word by part-of-speech) , then converted into a simpler WordNet tags
4. Words are Lemmatized, then assigned to WordNet Synsets
5. A word's Synset can be used to retrieve a sentiment score from the SentiWordNet lexicon
6. The scores for all of a tweet's words are summed. If the summed value is negative, a negative sentiment score is returned, otherwise a positive score is returned


Tweet Text
Cleaned up for analysis
Tokenization
simple tokens such as numbers, punctuation and words
Tagging (NLTK)
Role of each word
Convert to WordNet tags
Lemmatization
Remove inflectional endings (s, es, ing, ed, etc), return base form
WordNet Synsets
Sum SentiwordNet Scores
Return Positive or Negative rating

Weighted Engagement Scores: predicting the public sentiment of a movie

A core concept of this paper is *weighted engagement scores*, whereby the sentiment value of a given tweet may be given more weight towards the prediction of a movie's actual public sentiment based on the relative popularity of the tweet's author. In the data-mining phase, several attributes are extracted from each tweet towards this end: follower count, like count, and retweet count. An interesting preliminary finding was that on their own, these user engagement metrics had very little predictive value; a user with 1000 followers whose tweets receive no likes or retweets was not found to be more representative of public sentiment. Conversely, tweets where all 3 metrics had scores of 10 or greater were found to more closely correspond to public sentiment. Through an iterative process, the following algorithm was found to maximize the influence of user engagement scores:

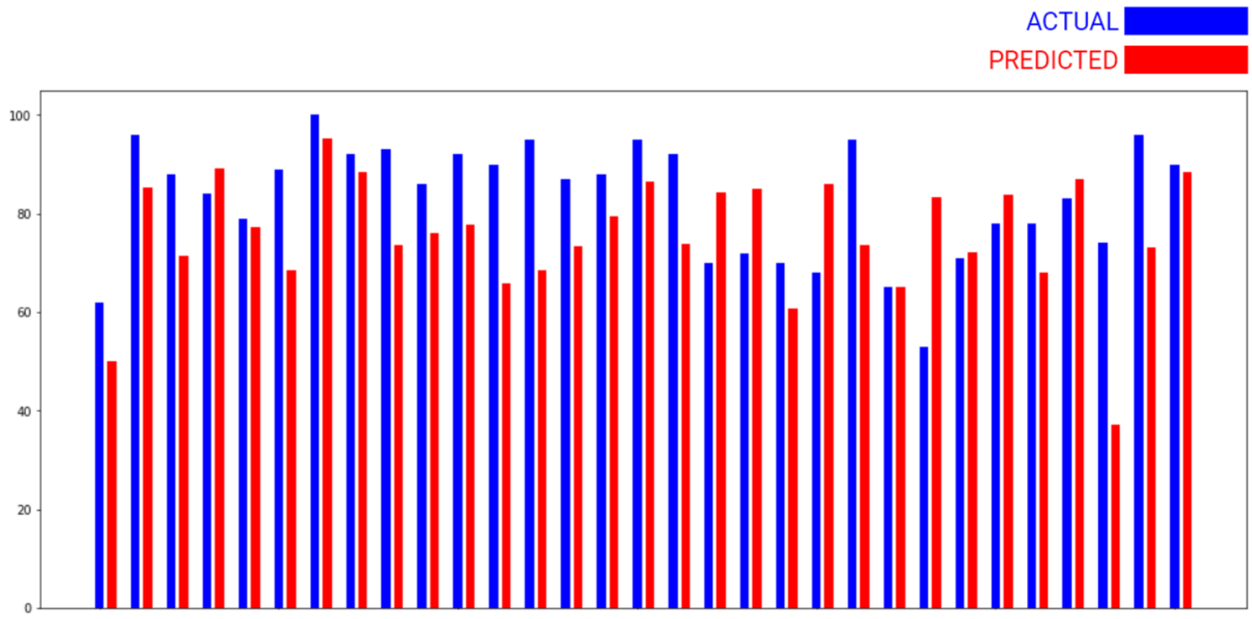


```
if(followers > 10 and likes > 10 and retweets > 10):
    weight = np.log(followers + likes + retweets)*1.5
```

Finally, a weighted average of scores is taken for all tweets about a given movie, then the movie is assigned a final score between 0 – 100 points.

RESULTS AND DISCUSSION

To gauge the accuracy of the algorithm on a given movie, it was necessary to find a source of data that was representative of public sentiment regarding movies. Audience Scores from the Rotten Tomatoes website filled this role nicely. Performance was calculated by direct comparison of the audience scores to the predicted scores on a scale from 0 to 100. The results can be seen below:

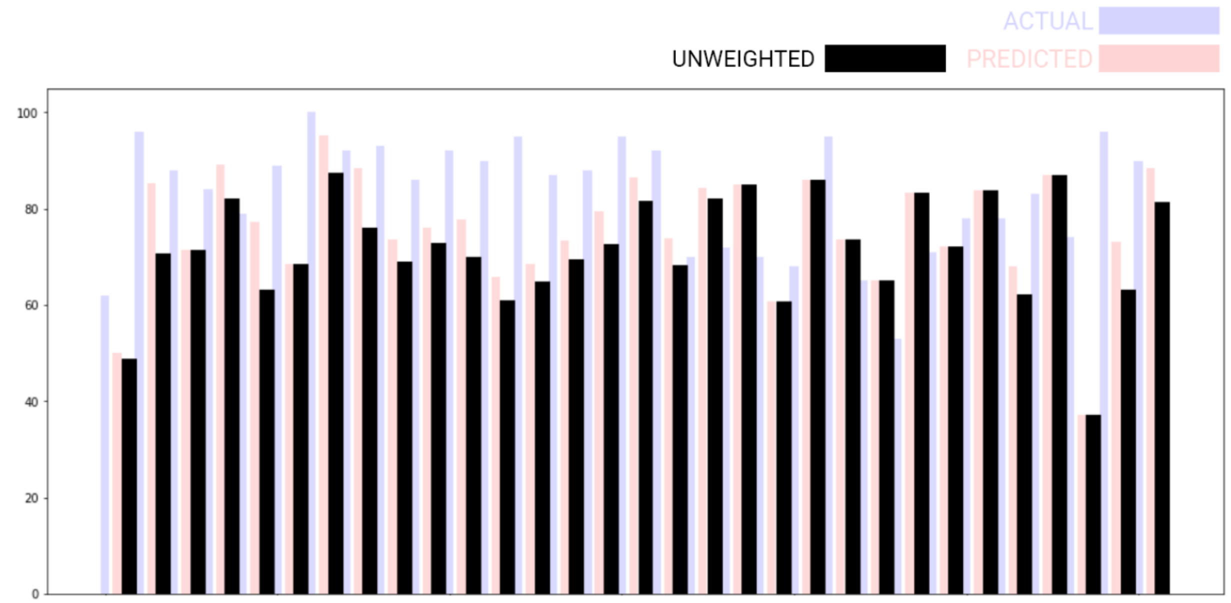


In general, predictions tracked actual values well, with only a few outliers. The overall measures of prediction error are shown below.

Percent Error	*16.3279*
Mean Absolute Error	13.1229

Weighted engagement scores

Weighted engagement scores were found to contribute a statistically significant improvement to overall scores, as demonstrated by the reduced accuracy of unweighted predictions seen below.



	Weighted	*unweighted*
Percent Error	16.3279	20.3371
Mean Absolute Error	13.1229	16.7439

Here we see that unweighted predictions performed consistently less well than the weighted predictions. It is worth noting that when the weighted and unweighted predictions have the same value, it indicates that there were no tweets for that movie that were sufficiently popular to pass the threshold for weighting. This tends to occur for movies that have relatively fewer tweets about them and suggests that further data mining for these movies may return better results.

With this trend in mind, future work will involve automating the process of tweet collection to build datasets large enough to support weighting, even for less popular films that tend to garner fewer tweets. The current system is designed to collect the maximum number of tweets for each film in succession, a process that takes a little over a 24 hours to complete. Programming it to run continuously for a week should collect sufficient tweets to analyze even smaller, independent films. Indeed, over 15 films from independent studios failed to meet the threshold for this paper including S*hooting the Mafia, Just Mercy*, and *Bean Pole*.

Future work will also focus on weeding out so called “opinion spammers”, automated bots posting tweets about films. Possible solutions include implementing spam detection, eliminating duplicate tweets, and exploring additional authority measures for twitter users.

CONCLUSION

Twitter provides a valuable resource for conducting sentiment analysis to accurately gauge the public's opinion of new movies. Not only does twitter offer a large volume of opinion-based text, but also provides useful metrics regarding the relative popularity of twitter users and their content. As shown in this paper, popular authors generally better reflect public sentiment.

Conducting Sentiment Analysis using Natural Language Toolkit and SentiWordNet proved to be an effective way to gauge the averaged public sentiment of a movie, but the predictions were not perfect. As weighted engagement scores contributed a 4% improvement in prediction accuracy, exploring ways to further refine the algorithm seems a promising place to begin.

REFERENCES

Ali Yadollahi, A. G. (2017). Current State of Text Sentiment Analysis from Opinion to Emotion Mining. *ACM Comput. Surv.*, Article 25, 33 pages.

Amolik, A. &. (2016). Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques. *International Journal of Engineering and Technology*, 2038-2044.

Andrew Perrin, M. A. (2019, April 10). *Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018*. Retrieved from Pew Research Center: https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/

Baid, P. e. (2017). Sentiment Analysis of Movie Reviews using Machine Learning Techniques. *International Journal of Computer Applications*, 45-49.

Chiplunkar, A. P. (2018). Real-time Twitter data analysis using Hadoop ecosystem. *Cogent Engineering*, 5:1.

Montoyo, A. &.-B. (2012). Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. *Decision Support Systems*, 675–679.

Pak, A. a. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *LREC*.

Shanahan, S.-H. K. (2003). Stigmatizing Smokers: Public Sentiment Toward Cigarette Smoking and Its Relationship to Smoking Behaviors. *Journal of Health Communication*, 343-367.