Twitter and Movies: Can Important Users Influence the Industry?

by

Jacob Mink

B.S. University of Vermont, 2010
B.A. University of Vermont, 2010

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment
of the requirements for the degree of
Master of Science
Department of Applied Mathematics
2017
This thesis entitled:
Twitter and Movies: Can Important Users Influence the Industry?
written by Jacob Mink
has been approved for the Department of Applied Mathematics

______________________________________
Dr. Brian Keegan

______________________________________
Dr. Jem Corcoran

______________________________________
Dr. Manuel Lladser

Date ______________________

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.
Over the past decade, social media has become an increasingly powerful tool for economic and social analysis. Specifically, Twitter has been used in a variety of fields to predict and analyze human behavior and the effects that behavior has on marketing and product sales. In this study, I attempt to find a predictive model for movie box office revenue using well-connected, film industry critics and other Twitter users, natural language processing, and statistical analysis.
Acknowledgements

I would like to thank my advisor Dr. Brian Keegan, and my thesis committee Drs. Jem Corcoran and Manuel Lladser.
## Contents

**Chapter**

1 Introduction .......................... 1

2 Literature Review ...................... 3
   2.1 Word of mouth importance .......... 3
   2.2 Analyze critics, not laypeople .... 4
   2.3 Twitter for box office prediction .. 5

3 Methods ................................ 7
   3.1 Data Acquisition .................. 7
   3.2 Natural Language Processing ...... 9
   3.3 Statistical Analysis ............... 11

4 Results ................................ 13

5 Discussion and Future Directions .... 24

Bibliography ............................. 26
Chapter 1

Introduction

Social media can be a powerful forecasting tool for certain subjects. Research done in the past about using Twitter to predict future events in the film industry has given mixed results in terms of the efficacy of Twitter as a predictive agent. In this study, I aim to analyze if using the appropriate dataset and Twitter user set, I can predict box office revenue of a set of films.

Why is Twitter a reasonable data set for analyzing critic sentiment effects on movie revenue? This question can be broken down into several user interaction categories. These categories are movie critics, movie industry people, and regular people.

The group of Twitter users I focus on in this study is comprised of well-connected users. These users utilize Twitter for various reasons. Non-industry users follow critics, actors, and other movie industry users to keep up-to-date on film events. The people they follow are the most likely to have first access to information about any entertainment news. People in the movie industry should follow other industry people for several reasons. In order to stay competitive in Hollywood, it makes sense for actors, directors, etc. to follow the Twitter feeds of other actors, directors, etc. Twitter is also useful in the film industry for one well-connected person to promote another, possibly less well-connected, user’s work in order to gain recognition for that person.

The argument for movie critics following movie industry people on Twitter is fairly self-explanatory, given the nature of their professional relationships. Twitter can be used
as a real-time data stream for critics, as well as a quotable source of entertainment news. The flow of information from actor, to Twitter, to movie critic retweeting that information is much more concise than the non-social media route of contacting agents and setting up interviews, etc.

The rationale for those in the movie industry following movie critics on Twitter could most likely be boiled down to a broad sense of egocentrism. If an actor wants to see themselves through the eyes of the public, a convenient way to do this would be to search for their own name in the tweets of movie critics. Additionally, actors could use critics’ tweets as an informal self-evaluation metric in order to improve their technique or manage the media’s view of them.

Increasingly over the past two decades, social media outlets have become important and vital sources of real-time, individualized public sentiment. Twitter is a prime example of a crowd-sourced data stream. With its short text format and simple user interface, it is easy for users to immediately publish their current thoughts and reactions, be it to politics, natural events, media news, or anything else they wish to publish. This contrasts wildly with traditional news and media formats, which have historically needed multiple phases from news event to publication. With platforms like Twitter, a professional movie critic, for example, is no longer restricted to their newspaper column and perhaps weekly publishing deadlines. On the contrary, they are now able to provide their views on any part of the film production process, from initial production announcement through box office release, whenever they want. This is a valuable resource for analyzing general sentiment about a film. In this project, I sought to crowdsourced and aggregate sentiments from a network of many movie critics and those Twitter users whom they follow. My thought was that this aggregation might tell me something about that movie, and specifically how well that movie was received.
Chapter 2

Literature Review

2.1 Word of mouth importance

The social phenomenon of "word-of-mouth" (WOM) has been studied across a variety of fields. With the development and growth of the Internet and online social media outlets, online WOM has become increasingly important for product reviews, sales, film and television approval ratings, and film box office revenue. Previous research has examined the effects of user reviews on film box office performance using several different online social networks. Some of the sites used are Yahoo! Movies ([5]; [8]; [14]; [16]), Variety.com ([8]), boxofficemojo.com ([8]; [15]; [17]), Rotten Tomatoes ([16]), imdb.com ([17]).

The research done in [5] involved finding relationships between online WOM and the level of success of "sequentially released new products". By this term they refer to theatrical releases of movies that have time lags between different regions. They study the effect of WOM reviews from early release regions on the subsequent ticket sales from later release regions.

In [5], the study finds several insights. First, they suggest that when early online WOM about a movie is negative, the production studio should delay that movie’s release in markets with low advertising presence and high WOM presence. Second, they suggest that when early WOM is positive, the studio should release the movie as soon as possible in those same types of markets. Thirdly, in markets with low advertising and low WOM presence, they suggest the studio should release the movie as late as possible in the release
cycle. These findings highlight the general importance of online communities to the success of a movie’s release.

2.2 Analyze critics, not laypeople

Studies have been done on the effects of film critics on movie success and the differences in effects between critic sentiment and non-critic (layperson) sentiment on movie success ([4]; [17]). In [4], they examine a random sample of both successful and unsuccessful films. For these films, the researchers analyzed weekly domestic revenue, valence of published reviews, a metric they cited from an earlier paper titled star power, budgets, and other control variables. Running regressions with all these variables on their dataset, the researchers found that the effects of critics’ reviews were significantly correlated with box office revenue, both positively for positive reviews and negatively for negative reviews. The researchers conclude that critics play a dual role of influencers and predictors. In other words, critics can not only reflect future viewer opinions in a passive way, but in fact influence those viewers and in turn affect a movie’s success.

The difference in criticism between professional critics, online critic communities, and random laypeople is studied in [17]. These researchers analyzed review data from metacritic.com, which is a repository for ”normalized, weighted average[s] of critics’ reviews from 42 major publications” ([17]). For online critic communities, data was obtained from both boxofficemojo.com and imdb.com. These critics were classified as ”self-defined” novices based on the fact that nothing about them was known save for their online movie critic presence in these communities. The third set of reviewers was a random sampling of 169 students at a university in the US that is not specified in the study. This study found significant variation in the way different types of people rate and review movies. They found that, in general, critics give lower ratings than novices, and those novices give lower ratings than laypeople. They speculate that many of the novice critics in online communities are in fact involved with the film industry in some capacity. This study highlights the differences in
review quality and analytical rigor of the three types of movie viewer. They also note that "Just as ... creativity is too diffuse to be constrained into ... labels, so too may expertise be unable to be pigeonholed into 'novices' and 'experts'." They suggest a continuum of movie expertise, on which critic, novice, and layperson are just three points of expertise.

2.3 Twitter for box office prediction

There has been research done on box office prediction using Twitter in the past. Although there have been mixed results in terms of finding correlation between Twitter activity and revenue prediction, these studies have presented a variety of methods for using Twitter data as a real-world analytical device in the film industry. Several of these studies utilize Twitter streaming data to attempt to forecast future events ([1]; [2]; [20]).

For [20], the researchers took the Twitter streaming data and converted each tweet into a feature vector of words and symbols. From this cleaned data, they classified the tweets about movies into positive or negative categories, and then compared this Twitter movie presence to user ratings on imdb.com and Rotten Tomatoes. The ultimate findings of this study were that there is a nonlinear relationship between high Twitter "hype" and high user ratings on other sites. However, they did find that if a movie has high Twitter hype and high user ratings, then it is likely to have a high box office revenue. With other combinatorics of Twitter hype and user ratings, however, the box office revenue range was fairly unpredictable.

The results from [20] conflict with earlier findings from [2]. In [2], the researchers gathered Twitter streaming data using movie titles as keywords, but they operated under a strong assumption that tweets containing those movie titles pre-release were assumed to be promotional about those movies. They also only considered tweets within a single week leading up to the theatrical release date. This study found a strong linear correlation between tweet counts within the week prior to release and first weekend box office revenue. This study is limited by the aforementioned assumption of Twitter presence having positive sentiment,
a small sample size of movies, and only accounting for the first weekend of revenue.

In [1], the researchers sought to summarize and expand upon previous studies using Twitter to predict box office revenue. Their methodology also relied on Twitter streaming data, with a dataset that encompassed films released from July 2011 to August 2011. They only utilized a small portion of this dataset for their analysis, and for this smaller dataset searched for correlation between tweet volume of tweets containing the movie title as keyword and box office revenue. Their findings suggest that there is causality between the daily number of tweets about a movie and that movie’s gross revenue. Another method they used was to define a sentiment index for the tweets in their dataset and search for correlations between these sentiments and box office revenue. Their findings were inconclusive regarding this relationship.
Chapter 3

Methods

Based on past research, I developed a central hypothesis to explore in this project.

**H1:** The buzz on Twitter from well-connected Twitter users in the movie industry should have an effect on the box office success of films mentioned.

The methodologies used to explore this hypothesis can be separated into three parts, namely data acquisition, natural language processing, and statistical analysis.

### 3.1 Data Acquisition

I chose the time range surrounding 2015 from which to get tweets, and the top 20 highest grossing films of 2015 to study. In order of highest-grossing first, these movies are: *Star Wars: The Force Awakens*, *Jurassic World*, *The Avengers: Age of Ultron*, *Inside Out*, *Furious 7*, *The Hunger Games: Mockingjay – Part 2*, *The Martian*, *Minions*, *Cinderella*, *Spectre*, *Mission: Impossible – Rogue Nation*, *Pitch Perfect 2*, *The Revenant*, *Ant-Man*, *Home*, *Hotel Transylvania 2*, *Fifty Shades of Grey*, *The Spongebob Movie: Sponge Out of Water*, *Straight Outta Compton*, and *San Andreas*. I chose to create three temporal eras for each of these films in which to find tweets. These eras are film-specific. For the first era, I arbitrarily chose a starting point of two years prior to the approximate US trailer release date, and an ending point of the trailer release date. The second era spans from the trailer release date to the theatrical release date for each movie. The third era spans from the theatrical release date to arbitrarily one year after that date. Henceforth, I will refer to
these eras as 'pre-trailer', 'post-trailer, pre-release', and 'post-release', respectively, or 'pre', 'post-pre', and 'post' for short.

In order to decide whose Twitter presence to analyze, I started with a seed set of Twitter users (9 well-known movie critics) that I obtained from the article “The Followables: 10 Film Critics You Should Follow on Twitter” (http://flavorwire.com/82825/the-followables-10-film-critics-you-should-follow-on-twitter). From this first ring of users, I then found all the users they follow, and then found all the users this second ring follows. This gave me a set of 1.6 million unique users. From this 1.6 million user pool, I narrowed the pool to only those members contained in both the second and third rings. The second ring of users contains an adequate number of users on which to base this project. In order to fully obtain the connectedness of every member of this ring, a third ring was needed. Taking the overlap of the second and third rings gives all interconnections within the second ring. This overlap contains exactly the same users as the second ring, but with more edges between these users.

The second and third ring of users are comprised of the following graphs of the previous ring’s users. I chose to use this sampling technique because I was searching for influential Twitter users. If an influential person follows another user, it is likely that that user is also influential, or at least important enough to be followed by someone of influence. Had I chosen to collect all the followers of the first ring of movie critics (instead of the followings), this would have given me a massive set of users that would not necessarily be well-connected in any way. For example, if I follow Roger Ebert’s Twitter account, my following him does not impart any increase in influence onto me. However, if Ebert’s account happened to follow me on Twitter, that would indeed impart some level of importance to me, simply because of Mr. Ebert’s societal influence. Once this initial sampling was done via this out-degree metric, I then switched to in-degree centrality for filtering the user set. Degree centrality gives a measure of connectedness for each node in a graph. It is defined as simply the number of links a certain node in a graph possesses. Degree centrality is comprised of in-degree and
out-degree directions, with in-degree being the number of links coming into a node, and out-degree being the number of links going away from a node. In a Twitter follower graph, this equates to in-degree denoting how many people are following the specified user, and out-degree denoting how many people the specified user is following. Therefore, in a study about influence, it is logical to examine each node’s in-degree centrality. I sorted the data by this metric, and selected the top 1000 most well-connected (highest in-degree centrality) users.

For each of the three film eras (pre, post-pre, post), I searched the Twitter users’ tweet history for the most recent 2000 tweets in those time periods. This tweet number limit was arbitrarily set to a high value in order to try to encompass all possible tweets from that user in each era. Once I had this corpus of tweets, I then reduced the corpus size by searching with a set of movie-specific keywords as filters. This set was comprised of the movie title seen as a single string, camel-capitalized with no spaces; the director’s first and last name as a single string; and the two top-billed actors’ first and last names. For example, the keyword list for the film *Pitch Perfect 2* is ['Elizabeth Banks', 'Anna Kendrick', 'Rebel Wilson', 'PitchPerfect2'].

### 3.2 Natural Language Processing

The second phase of my methodology was to apply natural language processing techniques to the acquired tweet text in order to glean some useful sentiment information. I utilized a Python package named VADER (Valence Aware Dictionary and sEntiment Reasoner), which is a sentiment analysis package specifically created for use on social media text (https://github.com/cjhutto/vaderSentiment). This analyzer is built to pick up on various sentiment hints in social media text, such as multiple capitalized letters in a row, the use of exclamation points, and the use of positive, negative, and “booster” words. These are words or phrases such as 'very', 'a lot', or 'uber'. Combined with other words in the text,
Figure 3.1: Revenue for each film
these "booster" words change the valence of the sentiment in the positive, negative, or neutral direction. VADER also outputs a composite sentiment score, which is a "normalized, weighted composite score" of the sentiment.

There are several sentiment analysis approaches available, and each method is useful in specific settings. For example, the lexicon LIWC analyzes the "emotional, cognitive, structural, and process components" of the input text sample. This method relies on a proprietary dictionary which is split into various categories, with 905 words comprising the two categories of obvious positive or negative emotions (Hutto, et al.). According to (Hutto, et al.), the shortcomings of LIWC in a social media context are that it does not include any method to analyze "acronyms, initialisms, emoticons, or slang", which are important and widely used on Twitter. Another way in which LIWC falls short for this project’s task is that it cannot measure the intensity of the sentiment-valued words in a text sample. Another text analysis tool which (Hutto, et al.) find lacking is the General Inquirer (GI), which has a larger lexicon including 1,915 positive-labeled terms and 2,291 negative-labeled terms. Even with this expanded lexicon, it has the same shortcomings as LIWC when applied to social media text. Based on the findings of (Hutto, et al.), I decided that VADER was the most applicable sentiment analyzer for this project’s content.

After running each individual tweet through VADER, I then weighted the sentiment scores by multiplying each tweet’s score by the in-degree centrality score of the user who produced it. These weighted scores were then sorted by movie and aggregated, giving a single positive, negative, neutral, and composite sentiment score for each era for each film. I then fit linear regressions to these sentiments vs. movie revenue. Each figure includes a lowess (locally weighted smoothing) plot for each movie era.

3.3 Statistical Analysis

The third phase of methodology was comprised of running various statistical analyses on the data I obtained. For this project, I only focused on single variable linear regression
models. Included within the data package for each tweet I pulled from Twitter was the number of mentions, retweets, date of tweet, and several other data types. In order to more fully analyze the tweets’ relationship to a film’s success, I searched for any correlation between the number of retweets of tweets about a specific movie and that movie’s box office revenue. I also compared average tweet text length and number of tweets per era to box office revenue.

I chose to utilize linear regressions with the initial data in order to test whether there was any simple correlations involved between the metrics obtained. These metrics were film revenue, retweet count, average tweet length, number of tweets, and sentiment analysis. Had I found some significant correlations using linear regression, I then would have expanded the statistical methodology to include more advanced techniques such as Random Forest or SVM (Support Vector Machine) algorithms. These would have been useful for increasing the precision and robustness of a working predictive model. However, at the stage of data analysis in which I was working, linear regressions were adequate for my needs.
Chapter 4

Results

The methodologies used in this project did not lead to many significant findings in correlating Twitter usage and film box office revenue. However, after running multinomial regressions for each movie era, it appears that the post.pre.retweet.count variable is significantly correlated with movie revenue. This might imply that the amount of times a tweet about a specific movie is retweeted between the time that of the trailer release and the theatrical box office release can influence how well that movie does at the box office.

As can be seen in Table 4.1, the top retweeted tweets from a sample of high-grossing movies have a variety of styles and subject matter. However, both the top retweeted tweets for Star Wars: The Force Awakens and The Avengers: Age of Ultron are promotional in content and highly positive in sentiment. The tweets for The Hunger Games: Mockingjay Part 2 both concern Jennifer Lawrence as an actress, and not the film directly, but when released during the post-pre era of this film could have possibly had a positive effect on the movie revenue.

From the various statistics I ran on the dataset, I found no obvious correlation between positive sentiment, average tweet length, or number of tweets and box office revenue. In Figure 4.1, the linear regressions have such high variances that not much can be concluded from these regressions. From Tables 4.2, 4.3, 4.4, and 4.5, these regressions individually do not have significant correlation, however Table 4.6 shows that when all these eras and total tweet counts are put into a multinomial regression with revenue, they produce a model that
<table>
<thead>
<tr>
<th>Movie</th>
<th>Number of retweets</th>
<th>Tweet text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Star Wars: The Force Awakens</td>
<td>31018</td>
<td>'Check out the official poster! Tune in to @ESPN 2019s Monday Night Football for a new look at TheForceAwakens . pic.twitter.com/NTRWLlVyID’</td>
</tr>
<tr>
<td>The Avengers: Age of Ultron</td>
<td>23644</td>
<td>'Fun way to reveal these, Marvel! Enjoy! Less than 2 months away! @Avengers AgeOfUltron pic.twitter.com/BDV1RbQFBX’</td>
</tr>
<tr>
<td>The Avengers: Age of Ultron</td>
<td>21774</td>
<td>'You’ve been good so here’s a new IronMan poster from @Avengers . &amp; on the DL, big announcement in 8 days getexcited pic.twitter.com/AdB1tTib7K’</td>
</tr>
<tr>
<td>The Hunger Games: Mockingjay Part 2</td>
<td>20265</td>
<td>'O Jennifer Lawrence I love you so. X’</td>
</tr>
<tr>
<td>The Martian</td>
<td>8444</td>
<td>'My favorite line in &quot;The Martian&quot; trailer, uttered by Matt Damon, is 'I’m going to have to science the shit out of this.’</td>
</tr>
<tr>
<td>The Martian</td>
<td>1201</td>
<td>'Andy Weir’s masterpiece! Jealous of Matt Damon! First in line for the movie? Me! Hanx pic.twitter.com/Ka7zTlyvWu’</td>
</tr>
</tbody>
</table>
has an $R^2$ value of 0.694. This denotes that the full model accounts for about 70% of the variance.

Table 4.2: Total tweet count vs. Revenue

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>Revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>total.tweet.count</td>
<td>11,783.680</td>
</tr>
<tr>
<td></td>
<td>(47,989.910)</td>
</tr>
<tr>
<td>Constant</td>
<td>278,953,073.000***</td>
</tr>
<tr>
<td></td>
<td>(56,085,822.000)</td>
</tr>
</tbody>
</table>

R$^2$ 0.003  
Adjusted R$^2$ -0.052 
F Statistic 0.060 (df = 1; 18)

*Note: $^*$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01
Figure 4.1: Various metrics vs. Revenue
Table 4.3: Pre-era vs. Revenue

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>Revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>pre.mean.tweet.length</td>
<td>$-11,740,867.000$</td>
</tr>
<tr>
<td></td>
<td>$(37,070,735.000)$</td>
</tr>
<tr>
<td>pre.tweet.count</td>
<td>$-247,176.800$</td>
</tr>
<tr>
<td></td>
<td>$(298,232.200)$</td>
</tr>
<tr>
<td>pre.retweet.count</td>
<td>$765.030$</td>
</tr>
<tr>
<td></td>
<td>$(1,265.029)$</td>
</tr>
<tr>
<td>Constant</td>
<td>$514,297,576.000$</td>
</tr>
<tr>
<td></td>
<td>$(646,890,530.000)$</td>
</tr>
</tbody>
</table>

$R^2 = 0.047$

Adjusted $R^2 = -0.132$

F Statistic $= 0.262$ (df = 3; 16)

Note: *p<0.1; **p<0.05; ***p<0.01
Table 4.4: Post-pre-era vs. Revenue

<table>
<thead>
<tr>
<th></th>
<th>Revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>post.pre.mean.tweet.length</td>
<td>$-15,389,330.000$</td>
</tr>
<tr>
<td></td>
<td>(32,079,875.000)</td>
</tr>
<tr>
<td>post.pre.tweet.count</td>
<td>$-103,074.800$</td>
</tr>
<tr>
<td></td>
<td>(138,105.000)</td>
</tr>
<tr>
<td>post.pre.retweet.count</td>
<td>$1,743.193^{***}$</td>
</tr>
<tr>
<td></td>
<td>(570.475)</td>
</tr>
<tr>
<td>Constant</td>
<td>$495,598,473.000$</td>
</tr>
<tr>
<td></td>
<td>(550,091,859.000)</td>
</tr>
</tbody>
</table>

$R^2$          0.468  
Adjusted $R^2$ 0.369  
F Statistic    $4.697^{**}$ (df = 3; 16)  

*Note:*  
*p<0.1; **p<0.05; ***p<0.01
Table 4.5: Post-era vs. Revenue

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>Revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>post.mean.tweet.length</td>
<td>7,222,350.000</td>
</tr>
<tr>
<td></td>
<td>(44,941,539.000)</td>
</tr>
<tr>
<td>post.tweet.count</td>
<td>84,445.730</td>
</tr>
<tr>
<td></td>
<td>(323,240.600)</td>
</tr>
<tr>
<td>post.retweet.count</td>
<td>-267.904</td>
</tr>
<tr>
<td></td>
<td>(679.799)</td>
</tr>
<tr>
<td>Constant</td>
<td>170,174,271.000</td>
</tr>
<tr>
<td></td>
<td>(750,643,579.000)</td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>R²</td>
<td>0.015</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>-0.169</td>
</tr>
<tr>
<td>F Statistic</td>
<td>0.083 (df = 3; 16)</td>
</tr>
</tbody>
</table>

*Note:* *p<0.1; **p<0.05; ***p<0.01
Table 4.6: All eras vs. Revenue

<table>
<thead>
<tr>
<th></th>
<th>Dependent variable:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Revenue</td>
</tr>
<tr>
<td>pre.mean.tweet.length</td>
<td>$-5,735,327.000$</td>
</tr>
<tr>
<td></td>
<td>(35,093,199.000)</td>
</tr>
<tr>
<td>post.pre.mean.tweet.length</td>
<td>$10,498,308.000$</td>
</tr>
<tr>
<td></td>
<td>(35,868,208.000)</td>
</tr>
<tr>
<td>post.mean.tweet.length</td>
<td>$23,407,378.000$</td>
</tr>
<tr>
<td></td>
<td>(40,785,289.000)</td>
</tr>
<tr>
<td>pre.tweet.count</td>
<td>$162,483.800$</td>
</tr>
<tr>
<td></td>
<td>(374,792.800)</td>
</tr>
<tr>
<td>post.pre.tweet.count</td>
<td>$-123,997.200$</td>
</tr>
<tr>
<td></td>
<td>(276,938.400)</td>
</tr>
<tr>
<td>post.tweet.count</td>
<td>$-182,447.500$</td>
</tr>
<tr>
<td></td>
<td>(346,074.000)</td>
</tr>
<tr>
<td>pre.retweet.count</td>
<td>$-1,462.273$</td>
</tr>
<tr>
<td></td>
<td>(1,946.368)</td>
</tr>
<tr>
<td>post.pre.retweet.count</td>
<td>$2,628.397^*$</td>
</tr>
<tr>
<td></td>
<td>(937.278)</td>
</tr>
<tr>
<td>post.retweet.count</td>
<td>$93.595$</td>
</tr>
<tr>
<td></td>
<td>(567.583)</td>
</tr>
<tr>
<td>total.tweet.count</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>$-213,923,326.000$</td>
</tr>
<tr>
<td></td>
<td>(692,278,081.000)</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.694</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.419</td>
</tr>
<tr>
<td>F Statistic</td>
<td>$2.520^*$ (df = 9; 10)</td>
</tr>
</tbody>
</table>

*Note:* $^*$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01
It is likely that the keywords I used to gather tweets related to the individual movies created a set of tweets that contained a lot of noise. For example, a tweet that contained the keyword “Michael Douglas” within the “pre-era” for the movie Ant-Man could easily be referring to another project or entertainment news story in which he was involved.

Additionally, I believe the VADER sentiment analyzer’s output for sentiment valence was incorrect in many cases and therefore did not give an accurate portrayal of the relationship between Twitter sentiment and box office revenue. These results are shown in Figure 4.2. From my hypothesis, I expected a strong linear correlation between positive sentiment and revenue. In Figure 4.2, the "Revenue vs. aggregate positive sentiment" plot shows some weak positive correlation for the post era, and a fairly decent positive correlation for the post-pre era starting at a sentiment score of about 0.015. The pre era shows a positive correlation up until about a sentiment score of 0.0125, but then shows a negative correlation. These plots lead me to believe that VADER is misclassifying many of the tweets’ sentiments.

While VADER was created to deal specifically with social media content, it falls short in recognizing certain important textual elements that would affect a tweet’s sentiment score. The fact that VADER is not designed to parse textual irony, nuance, or sarcasm from a piece of social media text certainly accounts for some of the misclassifications of tweets. VADER is designed to find sentiment clues in a body of text such as presence of exclamation points, number of capitalized letters, and certain “booster” vocabulary from a fixed set that denote either positive or negative sentiments. These features are not, however, necessarily going to occur in a tweet about a movie, even if that tweet contains overwhelmingly positive sentiment toward that movie. For example, if a user tweeted something like "I haven’t seen a movie this good in a very long time.", VADER would pick up on the word "good" and assign a slightly positive valence to the text. However, the overall sentiment of this example tweet should be classified as very positive. The user is expressing a very positive sentiment about the movie they have seen, but they do not use exclamation points or capitalized words to convey this sentiment. Since the high positivity of the text is not apparent from the various
cues that VADER can detect, it will thus misclassify the text. VADER is also unable to parse and interpret the presence of hashtags or URLs in a tweet, which probably caused it to misclassify a lot of text.
Figure 4.2: Weighted Aggregate Sentiment vs. Revenue
Chapter 5

Discussion and Future Directions

There has been some research done on sarcasm detection in product review text ([18]) that I believe can be tweaked to apply to Twitter-type text samples. In [18], Tsur uses Amazon book reviews as his dataset, but the methodologies can certainly be expanded to analyze social media text data. Unfortunately, when I contacted Tsur to ask if I could utilize his code on my data, he informed me that, due to circumstances beyond his control, he no longer has access to either the code or the original dataset for his paper. This code would be very useful in conjunction with VADER to overcome VADER’s lack of sarcasm and linguistic nuance detection. Future directions for this project would include attempting to develop this type of hybrid sentiment/sarcasm detection tool.

This project introduces several techniques for analyzing large-scale Twitter data that can be useful in a variety of settings, not only pertaining to movie industry applications. In future work, a more machine-learning based approach to keyword creation for searching tweet data will add to dataset precision and reduce noise in the form of unrelated tweets. I would like to be able to create a feature engineering scheme based on my seed keywords (actors, directors, movie titles) and expand that based on what other words are present in tweets that contain these keywords. From this, I may be able to improve the precision of my tweet-finding scheme by getting rid of tweets that do not involve the movie of interest.

Another direction for this project in the future could be to obtain each user’s full tweet history for the years 2014 and 2015, organize them by date, and run the keyword filters
again for each movie’s date range. Once this data is obtained, I will then get one tweet immediately preceding and one tweet immediately anteceding the tweet that contains the keyword. These dates could then be cross-referenced to each movie’s history/production cycle and any correlation between the tweets’ temporal location and real-world film industry events concerning the movie could be examined.

Yet another direction stemming from this project could be to find all the retweeted tweets in the dataset, and create stronger connections between those users that retweet each other’s content. This could help to separate the user network graph into film industry people, film critics, and normal users. Once this added layer of structure is applied to the graph, more important correlations between tweets and movie revenue could be found.
Bibliography


[3] Barthelemy, P., Guillory, D., Mandal, C. *Using Twitter Data to Predict Box Office Revenues*.


