Predict & Compare Success of Movies Based on Social Media

Raghavendra Prakash Nayak
Dept. of Computer Science
SUNY, University at Buffalo
Buffalo, New York 14260-2500
rnayak@buffalo.edu

Yashas Tiptur Ramesh
Dept. of Computer Science
SUNY, University at Buffalo
Buffalo, New York 14260-2500
yashasti@buffalo.edu

ABSTRACT

This paper introduces a new approach to predicting success of movies through opinion mining and web crawling. The paper explains the techniques used and the design details that went into the implementation of the predictive model and the system that we used to achieve our goal. To meet this goal, we computed the overall sentiment score that the movies received based on the twitter feed that we collected over a period of one week prior to the release of the movies. Machine learning techniques were used to compute the overall performance scores of the cast and the production house involved in the movie. A combination of the sentiment analysis and the performance scores was used to predict the success of these movies.

To test our approach, we compared our prediction results with results from Box Office Mojo. We found that our prediction was in-line to the results of Box Office Mojo. 7 out of 10 movies that we had predicted followed similar trends to that indicated in Box Office Mojo.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Miscellaneous;
H.3.3 [**Information Search and Retrieval**]: Miscellaneous

Keywords

Big Data, Text Mining, Predictive Analysis, Opinion Mining, Logistic Regression, Naïve Bayes Model

1. INTRODUCTION

Today, the social media had opened up a world where people create content, tag it, share it and network at a rate that is unprecedented. Examples include twitter, facebook, snapchat, pinterest etc. Every second, on average, around 6,000 tweets are tweeted on Twitter, which corresponds to over 350,000 tweets sent per minute, 500 million tweets per day and around 200 billion tweets per year. Because of the ease of use and wide spread reach of social media, it is fast changing the society and setting trends and agendas in topics ranging from environment to politics to technology and the film entertainment industry.

Since social media can be considered as a constructed form of collective wisdom, where we can investigate the chatter of a community in order to make quantitative decisions to predict the outcomes of products in those markets. These information markets generally involve the trading of state-contingent securities, and if large enough and properly designed, they are usually more accurate than other techniques for extracting diffuse information, such as surveys and opinions polls. Specifically, the prices in these markets have been shown to have strong correlations with observed outcome frequencies, and thus are good indicators of future outcomes. In this paper we predict the outcome in one such market – filmed entertainment industry.

Every production company tries to make the big decision of whether or not it should back a movie and in the recent past we have seen some major production houses lose enormous amounts of money backing the wrong projects. The consumers reaction to a possible new products has always been the focus of market research and has been rewarding, but in today's age can this be more accurate and yield realistic results. With films often costing $150 million and up, film studios need a better way than just intuition to ensure profits. The movie "John Carter" that released in 2012 (lost some $200 million for its investors) and it's easy to understand why. That's the reason many producers are moving beyond intuition and focus-group research and starting to turn to predictive analytics [1].

This paper is a study where we leverage the enormity and high variance of information from social media that propagates through large communities while presenting an opportunity for harnessing that data into a form that allows for predictions of particular outcomes. Our goal in this paper is to predict the outcome of movies before they are released. Our model also aggregates the information regarding past movie releases of the cast and production houses involved. This improves the accuracy while keeping in consideration the brand value that these entities bring to the overall predictions.

1.1 Problem Statement

To *predict* and compare the outcome of movies based on social media trends- twitter feed and reputation of the actors, directors, production houses based on all movies they have ever been a part of. Our model focuses on an extensive dataset mainly gathered from twitter and movie aggregator websites like IMDB, Rotten Tomatoes etc. This data will be used to analyze and predict the success of movies prior to their release.

1.2 Motivation

The movie industry is a multi-billion dollar industry and is highly volatile. It is difficult to say whether a big budget movie with a successful director and a super star can still rake in the profits. There have been many such examples of big budget movies that bombed in the box office one such movie being the Jhonny Depp starer "The Lone Ranger", this movie was based on a radio series that had a considerable fan base and involved an established crew. But the final outcome is difficult to predict and the solution to this problem has a very high significance.

We are trying to solve this very particular problem to some extent, by collecting data from focused sources and trying to draw parallels between social media chatter and box office success of the movies and then compare them to find the most successful of them. We will concentrate on movies that are similar in nature to avoid bias and increase the accuracy of our forecast. The system will basically work on the viewer's excitement over the release of a movie and the reputation of the cast and the production houses involved in the development of the movie.

2. RELATED WORK

Twitter has been a very popular web service, it has become almost omnipresent with hyper-connectivity for social networking and content sharing. The website www.internetlivestats.com, a live tweet counting website, has calculated that every second, on average, 6000 tweets are tweeted on Twitter which corresponds to over 350,000 tweets sent per minute, 500 million tweets per day and around 200 billion tweets per year [2].

This brings us to people, and how they share their views. Nowadays, a large number of people raise their voice via social media and this has led to emergence of companies that provide analytics by using the social media. Therefore, there has been certain work done in the field of predictive analytics using Twitter and other social media platforms.

Some of the works include predicting revenues that the movie might make based on reviews [3], where the system predicts the revenue that would be generated based on the reviews collected from movie aggregator websites such as BoxOfficeMojo, IMDb etc.… after the movie has released. There is also some work done in early prediction of movie success based on features such as “who”, “what” etc.… obtained from the same aggregator websites and running machine learning algorithms to predict the success of the movie [4]. There is also work done in predicting the success of a movie using sentiment analysis of twitter feeds and YouTube comments [5]. However, this has been done to predict the success of only one movie. Our model uses a novel technique to predict the success of movies, we don’t just assess the sentiment scores of the movies, but also consider the reputation and performance scores of the cast and production house will predicting the final outcome.

3. SYSTEM OVERVIEW

The figure-1 represents the framework of our predictive model. The first phase is data acquisition, because our prediction is based on twitter feeds and the data available from Rotten Tomatoes and IMDB. We will be using both these datasets in our system.

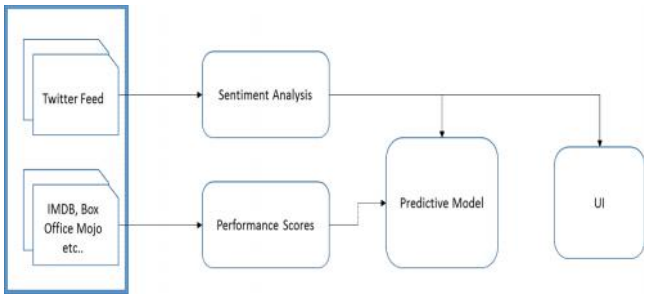

Figure-1

In the first phase of data acquisition, we collected over 150K tweets from twitter for 10 movies over a period of one week. We however, have not considered the genre of the movies. This was done because, in our model we are not comparing the performance of the movies against one another, but we are comparing the outcome of our prediction when compared to the real world results. We also scrapped the movie aggregator websites to generate a database of all the movies that the cast or production house was ever involved in. This was done to generate the performance scores of the cast and production house based on past reputation. A dataset of almost 1000 movies was generated and classified.

In the second phase, sentiment classifier was trained on a polarity corpus of Pang Lee consisting of over 10000 positive and negative short reviews. Additional tweets were added to the corpus and the classifier was trained again to increase the accuracy of the sentiment analysis. We used logistic regression to compute the overall performance score of the movie and we also computed the individual performance scores of the actors and directors involved in the movie by aggregating the scores for each of their previous movies. The overall performance score and the overall sentiment score was then used to predict and compare the success of movies.

4. DATASET

4.1 Twitter Dataset:

In order to accurately make a prediction of the outcome of movies, we needed a large dataset that we collected from twitter. The dataset from twitter was obtained by crawling twitter for each of the movies for a fixed duration of time. To ensure that we obtained all tweets referring to a movie, we used keywords present in the movie titles, the lead cast and directors of the movie along with the movie name as search arguments. Care was taken not to include any other movie than the one being queried for. Also, additional care was taken not to include any advertisement tweets of the movies. This was so avoid unfair advantages that the

movie might get while computing the sentiment scores. Also, advertisements can't really be considered as sentiments of twitter users.

We extracted tweets every day using the *Twitter Search Api*, ensuring we had the twitter handle name, time stamp, time zone, tweet text and the user id for our analysis. Time zone was used to locate the origin of tweet and visualize it graphically. We collected approximately over 150,000 tweets for 10 different movies over a period of one week.

Movies are typically released on Fridays, with exception of few that are released on Wednesdays. Also, since there is a lot of social media chatter in the week the movie is set to release, we tried to exploit this by collecting the tweets for a period of one week till one day prior of the movie's release date. Some details on the movies chosen and their release dates are provided in Table.1

Movie	Release Date
The Huntsman: Winter's War	04/22/2016
The Jungle Book	04/15/2016
The Barbershop: Next Cut	04/15/2016
Keanu	04/29/2016
Ratchet & Clank	04/29/2016
Mother's Day	04/29/2016
The Family Fang	04/29/2016
The Man Who Knew Infinity	04/29/2016
Pele: Birth of a Legend	05/06/2016
Captain America: Civil War	05/06/2016

TABLE I
NAMES AND RELEASE DATES OF MOVIES WE CONSIDERED IN OUR ANALYSIS

The tweet volume that we crawled has been shown in the Table 2. and Figure I. The figure shows a fairly uneven tweet volume distribution.

Movie	Tweet Volume
The Huntsman: Winter's War	15561
The Jungle Book	95445
The Barbershop: Next Cut	3050
Keanu	697
Ratchet & Clank	1641
Mother's Day	1135
The Family Fang	660
The Man Who Knew Infinity	400
Pele: Birth of a Legend	630
Captain America: Civil War	83780

TABLE II
NAMES AND TWEET VOLUME OF MOVIES WE CONSIDERED IN OUR ANALYSIS

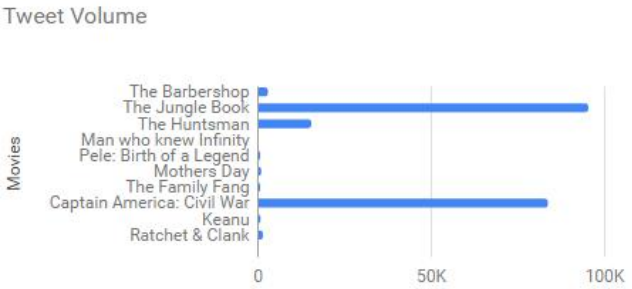

FIGURE I
TWEET DISTRIBUTION ACROSS MOVIES

4.2 Movie- Dataset:

We wanted to generate a dataset that takes into consideration the track record of the actors and directors that have acted in the movie whose success we want to predict. Hence we created this data set that consists of all the movies that an actor has acted in or a director has directed. For example one of the movies whose success we want to predict is "The Jungle Book" and one of the leading actors in the movie is Sir Ben Kingsly, he has acted in about 130 movies so we consider all these movies in our data set. For directors who have acted in as well as directed the movie we considered the previous movies that they have acted in and the movies they have directed separately, their track record as actors and as directors were considered independently. In a similar way we collected this data for the three main actors from each of the ten movies that we considered in our system. We considered the Production House, IMDB_Rating, IMDB_Votes, Metascore, Tomato_Meter, Tomato_Rating, Tomato_Review, Tomato_Fresh, Tomato_Rotten, Tomato_User_Rating, Tomato_User_Review, Budget to create our data set.

The data set includes data scraped from movie aggregator websites IMDb, Rotten Tomatoes and Metacritic which have their own API's like OMDb API and Rotten Tomatoes API etc. From these websites we collected information about the movies IMDb rating, number of votes on IMDb, metascore, tomato meter, tomato user rating, tomato reviews, votes for 'want to watch' and other scores. We considered 21 such scores, this is seen in the Figure II shown below;

Type	Features
Nominal	Actors, Director, Writer, Production-House,
Numerical	Budget, IMDB Rating, No of Rating, IMDB Votes, Metascore, Tomato Meter, Tomato User Rating, Tomato Reviews, Tomato Fresh, Tomato Rotten.

FIGURE II
FEATURE SET OF ALL PREVIOUS MOVIES THE CAST HAS BEEN A PART OF

The values that were scraped from IMDb were the director rating, actor rating, Production Company, IMDb votes and metascore. Metascore which is the weighted average of the critics rating that is published by the Metacritic website they compile this score by assigning more importance, or weight, to some critics and publications than others, based on their quality and overall stature then they normalize the resulting scores. We gave a higher weight to this score. We used the IMDb API to scrape these values.

Tomatometer, Tomato rating, Tomato reviews, Tomato fresh, Tomato rotten and Tomato user review these were the scores that we collected from the Rotten Tomatoes website, each of the scores indicates different values that were considered like the Tomato reviews and Tomato user reviews were the consolidation of the number of reviews that the movie received from critics as well as from the users of the web site, the user votes based on whether they wanted to watch the movie or not were also considered and based on the viewers response the movies were marked as fresh or rotten. The budget of the movie was also considered

Once these performance scores were calculated they were modeled so that they can be handled by our machine learning model. The main data set consists of data for approximately 1566 movies.

5. DESIGN & IMPLEMENTATION:

5.1 Sentiment Analysis:

We used NLTK and Naïve Bayes Classifier to implement our Sentiment Classifier. The construction of the same has been explained below.

Data Preprocessing:

Data preprocessing consists of three steps: 1) tokenization, 2) normalization, and 3) part-of-speech (POS) tagging. For the normalization process, the presence of abbreviations within a tweet is noted and then abbreviations are replaced by their actual meaning (e.g., BRB − > be right back). We also identify informal intensifiers such as all-caps (e.g., THE JUNGLE BOOK rocks!!! and character repetitions (e.g., I've love Batman!! happyyyyyy"), note their presence in the tweet. All-caps words are made into lower case, and instances of repeated characters are replaced by a single character. Finally, the presence of any special Twitter tokens is noted (e.g., #hashtags, usertags, and URLs) and placeholders indicating the token type are substituted. Our hope is that this normalization improves the performance of the POS tagger, which is the last preprocessing step.

Part-of-Speech Tagging:

For each tweet, we have features for counts of the number of verbs, adverbs, adjectives, nouns, and any other parts of speech. We have used the Stanford **N**atural **L**anguage Processing **T**ool **K**it's (NLTK) POS tagger. Part-of-speech tagger is a piece of software that reads text and assigns part of speech to each word such as adjectives, nouns etc. In our sentiment analysis model, we have considered the adjectives in classification of positive and negative polarity sets.

Naïve Bayes Classifier

A Naive Bayes Classifier is a simple probabilistic model based on the Bayes rule along with a strong independence assumption.

The Naïve Bayes model involves a simplifying conditional independence assumption. That is given a class (positive or negative), the words are conditionally independent of each other. This assumption does not affect the accuracy in text classification by much but makes really fast classification algorithms applicable for the problem. Rennie et al discuss the performance of Naïve Bayes on text classification tasks in their 2003 paper.

In our case, the maximum likelihood probability of a word belonging to a particular class is given by the expression:

$$P(x_i|\ c) = \frac{Count\ of\ x_i\ in\ documents\ of\ class\ c}{Total\ no\ of\ words\ in\ documents\ of\ class\ c}$$

The frequency counts of the words are stored in hash tables during the training phase. According to the Bayes Rule, the probability of a particular document belonging to a class ci is given by,

$$P(c_i|d) = \frac{P(d\ |c_i) * P(c_i)}{P(d)}$$

If we use the simplifying conditional independence assumption, that given a class (positive or negative), the words are conditionally independent of each other. Due to this simplifying assumption the model is termed as "naïve".

$$P(c_i|d) = \frac{(\prod P(x_i|c_j)) * P(c_j)}{P(d)}$$

Here the xi s are the individual words of the document. The classifier outputs the class with the maximum posterior probability. We also remove duplicate words from the document, they don't add any additional information; this type of Naïve Baye algorithm is called Bernoulli Naïve Bayes. Including just the presence of a word instead of the count has been found to improve performance marginally, when there is a large number of training examples.

To initially train our sentiment classifier, we used the Pang/Lee polarity corpus of short reviews from NLTK. The corpus contains 5331 positive and 5331 negative sentences/snippets. This dataset was introduced in 2005 by Pang/Lee. However, we also added tweets that were manually annotated by us into this dataset. This was done to increase the accuracy of the classifier.

5.2 Performance Scores:

We used Logistic regression to generate our performance score for each of the movies. The construction of this model is as follows.

Actors and Directors Rating:

To create an extensive data set on which we ran our regression model we needed a score that indicated the actors or directors performance in their previous movies. We wanted to consider the track record of the cast of the movie while predicting the success of the movie. To achieve this we considered the actors and directors in a movie and all the movies that they have been a part of in the past and the success of those movies. For each actor/director that we considered, we first obtained each of their previous movies' IMDb rating and metascore, we then aggregated the value of all these scores and computed a score or value between 0 and 10. This score indicated the individual actor's/director's performance score based on their track record. This value was then used in combination with the other values that we scraped.

For example, for the movie The Jungle Book, we calculated the actors and director's performance score as shown in Table III-

Actor/Director	**Performance Score (Range – 0 to 10)**
Jon Favreau	7.2
Ben Kingsly	7.9
Bill Murray	6.6
Idris Elba	6.8

TABLE III
PERFORMANCE SCORES OF ACTORS AND DIRECTOR OF THE JUNGLE BOOK

These scores were then used to find the overall performance of the movies.

Logistic regression:

For this project we considered ten movies whose success we predicted. To achieve this the data set that we created consisted of all the movies that the actors and directors in the ten movies that we have considered acted in. For each movie apart from the various IMDb scores, Metacritic scores and the Rotten Tomatoes score we also considered the actors and directors rating that we generated as explained above. Our system uses logistic regression to predict an overall performance score for each of the ten movies that we have selected based on the data set that we created.

Logistic regression, is a linear model for classification rather than regression. In this model, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function with the equation as follows:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

We then used IMDb, Rotten Tomatoes and Metacritic to scrape ratings and votes for each of the ten movies that we are considering. We considered the following values for each movie Director, Rating, Writer, Actor1, Actor2, Actor3, Production, IMDB_Rating, IMDB_Votes, Metascore, Tomato_Meter, Tomato_Rating, Tomato_Review, Tomato_Fresh, Tomato_Rotten, Tomato_User_Rating, Tomato_User_Review, Budget. This was used to create a dataset which ranged up to 1566 movies.

The logistic regression model was trained using this data set of 1566 movies. We then used this trained model to classify the 10 movies selected into a range of 0-3, where 0 indicated that the movie would perform poorly and 3 indicated that the movie would be a blockbuster. These results were then used in the final prediction model along with the sentiment scores to predict the success of movies.

5.3 Prediction Model

The final predictive model is a combination of sentiment analysis and the overall performance score that was generated for the movie based on all the movies that the cast were ever a part of.

The overall sentiment score was calculated for each of the 10 movies. We used the P/N ratio method to aggregate the total positive and negative tweets received by the movie. P/N ratio is a technique to calculate average sentiment score across multiple items. The technique is by Lexalytics Inc. an analytics company based out of Boston, USA. The pseudo-code for the P/N ratio technique is given below:

IF pos_count < neg_count THEN
Ratio = -1 x (neg_count / pos_count)
IF pos_count > neg_count THEN
Ratio = 1 x (pos_count / neg_count)
ELSE Ratio =0

IF ratio < -10 THEN ratio = -10
IF ratio > 10 THEN ratio = 10

The final sentiment score that was generated using the pseudo-code is given in the below Table IV.

Movie	**Sentiment Score**
The Huntsman: Winter's War	-2.49
The Jungle Book	1.57
The Barbershop: Next Cut	-1.63
Keanu	1.46
Ratchet & Clank	-1.84
Mother's Day	-1.825
The Family Fang	-10
The Man Who Knew Infinity	0
Pele: Birth of a Legend	-1.52
Captain America: Civil War	1.23

TABLE IV
SENTIMENT SCORE GENERATED FOR THE MOVIES

This score was combined with the overall performance score which was normalized to a range of 0-10.

Prediction Score = Sentiment_Score(0.3) + Overall Performance Score

Thus the final score – a result of sentiment analysis and performance scores was as below in Table V;

Movie	Prediction Score
The Huntsman: Winter's War	3.6
The Jungle Book	7.6
The Barbershop: Next Cut	4.2
Keanu	6.5
Ratchet & Clank	4.2
Mother's Day	4.7
The Family Fang	1.1
The Man Who Knew Infinity	6.3
Pele: Birth of a Legend	4.7
Captain America: Civil War	6.8

TABLE V
FINAL PREDICTION SCORE

6. RESULTS

Sentiment Analysis:

The accuracy of the sentiment classifier when trained on the Pang/Lee dataset of 10000 short reviews is given below – here the training set consisted of 80% of the corpus and the testing set was the rest 20% of it. We obtained an accuracy of 77.8%, it has been shown in the screen shot of python terminal shown below:

```
Evaluation performed on NLTKPang/Lee Movie Corpus - 10000 reviews
Evaluation of the training corpus - for Naive Bayes Classifier
Trained on 8528 Tweets, Tested on 2134Tweets
Sentiment Analysis (Bayesian Classifier) Accuracy: 0.778819119025
Pos Tweets - Precision: 0.787996127783
Neg Tweets - Precision: 0.770208900999
Pos Tweets - Recall: 0.762886597938
Neg Tweets - Recall: 0.794751640112
Pos Tweets - Fmeasure: 0.775238095238
Neg Tweets - Fmeasure: 0.782207022070
```

FIGURE III
ACCURACY OF SENTIMENT ANALYSIS CLASSIFIER

The accuracy and f-measure of the sentiment classifier was computed for each of the 10 movies we had chosen. This has been shown in the Table VI & Table VII shown below;

Accuracy:

Movie	Accuracy
The Jungle Book	73.07%
The Huntsman: Winters War	56.10%
The Barbershop: Next Cut	68.22%
Ratchet & Clank	67.34%
Keanu	73.41%
Mother's Day	62.25%
High-Rise	60.00%
The Family Fang	54.70%
Pele: Birth of a Legend	60.31%

f-measure:

Movie	f-measure	
	Pos Tweets	Neg Tweets
The Jungle Book	0.717	0.675
The Huntsman: Winters War	0.411	0.526
The Barbershop: Next Cut	0.622	0.562
Ratchet & Clank	0.512	0.615
Keanu	0.783	0.655
Mother's Day	0.445	0.397
High-Rise	0.615	0.285
The Family Fang	0.394	0.473
Pele: Birth of a Legend	0.561	0.637

TABLE VII
F-MEASURE RESULTS OF THE SENTIMENT CLASSIFIER

The sentiment separation – i.e. positive sentiments and negative sentiments collected over a defined time period is shown in the below Figure IV. Here the positive sentiments are visualized in blue and the negative sentiments are in red.

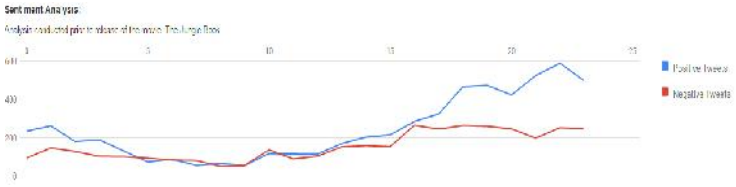

FIGURE IV
TREND OF POSITIVE & NEGATIVE SENTIMENTS FOR THE JUNGLE BOOK

Performance Score

The result of our regression model was dependent on the size of the training sample. When we increased the sample size the accuracy also increased. It was also observed that the accuracy was dependent on whether the samples were randomized, for random training data samples the accuracy was higher. When we used 80% of the data set as the training data we achieved an accuracy of 61.2% which was the highest accuracy achieved. This is shown in the below Figure V.

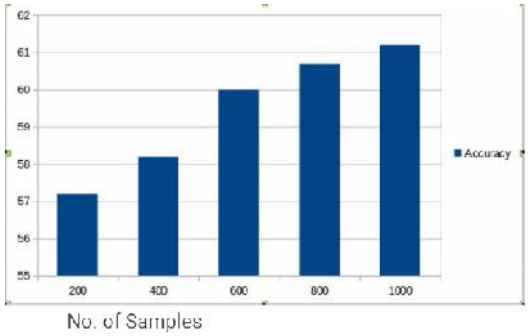

FIGURE V
ACCURACY OF LOGISTIC REGRESSION BASED ON NO. OF SAMPLES

Evaluation of Prediction Results:

Our model does not exactly predict the gross collection the movie made or its Box Office Collection, we have used the real world ratings for the movies as comparison to evaluate how accurate our prediction model is.

We can observe from our results that using only sentiment analysis or a performance score we cannot accurately predict the success of a movie. A combination of the makes it more likely for the prediction to correct. When we compared our results with the real world outcome we observed that we got 7 out of the top 10 movies featured on Box Office Mojo correct. The comparison is made with the first weekend box office results of the movies obtained from Box Office Mojo.

We observed that our prediction model accurately predicted that "The Jungle Book" would be the most successful movie in the weekend box office. We also predicted that the movie "Keanu" would also be successful and rated it second right after "The Jungle Book" which was in line with the real world results, so was our prediction for "Mother's Day" and "Ratchet and Clank". But our prediction that "The Huntsman: Winter's War" did not perform well was incorrect and was not the same as the real world results, the movie was placed third in the weekend box office ratings.

These results have been shown in the Table VIII below;

Movie	Prediction Score	Prediction Rank	Box Office Mojo
The Huntsman: Winter's War	3.6	7	3
The Jungle Book	7.6	1	1
The Barbershop: Next Cut	4.2	6	5
Keanu	6.5	2	2
Ratchet & Clank	4.2	6	7
Mother's Day	4.7	5	4
The Family Fang	1.1	8	34
The Man Who Knew Infinity	6.3	4	28
Pele: Birth of a Legend	4.7	5	Not released
Captain America: Civil War	6.8	3	Not released

TABLE VIII
COMPARISION OF PREDICTION & RESULTS FROM BOX OFFICE MOJO

The table shows that there is some correlation between the results from Box Office Mojo and our predicted results. This is only a comparison of how the movies will perform against one another in the box office. We can see that The Jungle Book performed the best amongst all the other movies where as The Family Fang performed the worst. Also, we can see that Mother's Day and Barbershop performed equally well in the box office.

7. REFERENCES

[1]. How Predictive Analysis is Changing Hollywood – Business Insider http://www.businessinsider.com/sc/predictive-analytics-is-changing-hollywood-2013-8

[2]. Twitter Usage Statistics – Internet Live Stats http://www.businessinsider.com/sc/predictive-analytics-is-changing-hollywood-2013-8

[3]. Mahesh Joshi Dipanjan Das Kevin Gimpel Noah A. Smith. Movie Reviews and Revenues: An Experiment in Text Regression *Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies Conference, Los Angeles, CA 2010*

[4]. Michael T. Lash and Kang Zhao. Early Predictions of Movie Success: the Who, What, and When of Profitability. *Submitted on 17 Jun 2015* http://arxiv.org/abs/1506.05382v2

[5]. Yafeng Lu, Robert Kruger, ¨ Student Member, IEEE, Dennis Thom, Feng Wang, Steffen Koch, Member, IEEE, Thomas Ertl, Member, IEEE, and Ross Maciejewski, Member, IEEE. Integrating Predictive Analytics and Social Media. *Published in Visual Analytics Science and Technology (VAST), 2014 IEEE Conference.*

[6]. M. Saraee, S. White & J. Eccleston. A data mining approach to analysis and prediction of movie ratings, *Fifth International Conference on Data Mining, Text Mining and their Business Applications,, 15-17 September 2004.*

[7]. Sitaram Asur and Bernardo A. Huberman. Predicting the Future With Social Media, March 2010. http://arxiv.org/abs/1003.5699v1

[8]. Kazem Jahanbakhsh, Yumi Moon. The Predictive Power of Social Media: On the Predictability of US Presidential Elections using Twitter. *Submitted on July 2014.* http://arxiv.org/abs/1407.0622v1

[9]. Stefan Nann, Krauss Jonas, Schoder Detlef. Predictive Analytics on Public Data- The Case of Stock Markets. *ECIS 2013 Completed Research. Paper 102.* http://aisel.aisnet.org/ecis2013_cr/102

[10]. IMDb Database Statistics, 2014. http://www.imdb.com/stats

[11].Natural Language Tool Kit. http://www.nltk.org/

[12]. Twitter API documentation. https://dev.twitter.com/overview/documentation

[13]. Rotten Tomatoes API documentation. http://developer.rottentomatoes.com/

[14]. Box Office Mojo Feeds. http://www.boxofficemojo.com/about/data.htm

[15]. Bing Liu. Sentiment Analysis and Opinion Mining.

[16]. Weekend Box Office Results – Box Office Mojo http://www.boxofficemojo.com/weekend/chart/?yr=2016&wknd=18&p=.htm

[17]. Sitaram Asur and Bernardo Huberman – 'Predicting the Future with Social Media – HP Labs. *Submitted on 29 March 2010* https://arxiv.org/pdf/1003.5699.pdf

[18] Jure Leskovec, Lada A. Adamic and Bernardo A. Huberman. The dynamics of viral marketing. In Proceedings of the 7th ACM Conference on Electronic Commerce, 2006.

[19] Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. First Monday, 14(1), Jan 2009. http://arxiv.org/pdf/0812.1045.pdf

[20] B. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology, 2009. http://ir.cs.georgetown.edu/publications/downloads/Twitter_power-_Tweets_as_electronic_word_of_mouth.pdf

[21] D. M. Pennock, S. Lawrence, C. L. Giles, and F. A. Nielsen. The real ˚ power of artificial markets. Science, 291(5506):987 – 988, Jan 2001. https://clgiles.ist.psu.edu/papers/Artificial_Markets_Science01.pdf

[22] Kay-Yut Chen, Leslie R. Fine and Bernardo A. Huberman. Predicting the Future. Information Systems Frontiers, 5(1):47–61, 2003. http://www.eecs.harvard.edu/cs286r/courses/fall12/papers/CFH03.pdf

[23] W. Zhang and S. Skiena. Improving movie gross prediction through news analysis. In Web Intelligence, pages 301304, 2009. http://www.cs.cmu.edu/~nasmith/TDF/ZhangWenbinISF2009Paper.pdf

[24] Akshay Java, Xiaodan Song, Tim Finin and Belle Tseng. Why we twitter: understanding microblogging usage and communities. Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, pages 56–65, 2007. http://ebiquity.umbc.edu/paper/html/id/367/Why-We-Twitter-Understanding-Microblogging-Usage-and-Communities

[25] Ramesh Sharda and Dursun Delen. Predicting box-office success of motion pictures with neural networks. Expert Systems with Applications, vol 30, pp 243–254, 2006. http://cs229.stanford.edu/proj2011/ImNguyen-PredictingBoxOfficeSuccess.pdf

[26] Daniel Gruhl, R. Guha, Ravi Kumar, Jasmine Novak and Andrew Tomkins. The predictive power of online chatter. SIGKDD Conference on Knowledge Discovery and Data Mining, 2005. http://dl.acm.org/citation.cfm?id=1081883

[27] Mahesh Joshi, Dipanjan Das, Kevin Gimpel and Noah A. Smith. Movie Reviews and Revenues: An Experiment in Text Regression NAACL-HLT, 2010.

[28] Rion Snow, Brendan O'Connor, Daniel Jurafsky and Andrew Y. Ng. Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. Proceedings of EMNLP, 2008.

[29] Bo Pang and Lillian Lee. Opinion Mining and Sentiment Analysis Foundations and Trends in Information Retrieval, 2(1-2), pp. 1135, 2008.

[30] Namrata Godbole, Manjunath Srinivasaiah and Steven Skiena. LargeScale Sentiment Analysis for News and Blogs. Proc. Int. Conf. Weblogs and Social Media (ICWSM), 2007. http://www.icwsm.org/papers/3--Godbole-Srinivasaiah-Skiena.pdf