Predicting Movie Statistics Using Twitter

Aditi Singh
Department of Computer Science
Stony Brook University
adisingh@cs.stonybrook.edu

Rahul Rishi Sharma
Department of Computer Science
Stony Brook University
rrsharma@cs.stonybrook.edu

Tanya Tiwari
Department of Computer Science
Stony Brook University
ttiwari@cs.stonybrook.edu

ABSTRACT

Social Media has been emerging over the years and so is the data associated with it. Websites such as Twitter and Facebook generate large amount of data in the form of tweets and posts, which can be useful in predicting the future. In this project we aim at predicting the movie revenues by harnessing tweets collected from Twitter. We collect real time tweets using twitter API and also perform sentiment analysis to review a movie's success. Finally, we develop a linear model to predict the revenue of the movies given certain movie features (e.g. tweet count, tweet polarity, user influence etc).

Keywords

Twitter; Movie Revenue

1. INTRODUCTION

The average person spends around 1 hour and 40 minutes on social media every day, accounting for 28 % of the total time spent on the internet.This is one of the main reasons for the large amount of data available on these websites.Users usually generate content, access information to reach a large audience. Social media has replaced the traditional one-way mass media to consumer communication channel with an interactive dialogue, which helps in creation and exchange of user-generated content. Twitter, a social media website, has a huge user base responsible for generating huge volumes of rapidly changing data content.

Our main goal in this project is to analyze tweets ,to eventually build a model which is able to predict box office revenues. The project is divided into three phases.

- Data Collection
- EDA on the data collected
- Using regression model and sentimental analysis build a model to predict box office revenues.

2. RELEVANT WORK

Using social media to be able to predict future trends in various fields is generating a lot of interests among researchers and data scientists.[1] tries to show how sentiment analysis on tweets is used to predict the success of a movie. [2] uses twitter and YouTube data to predict the IMDB scores of movies.[3] shows how social media helps to predict the future by doing regression analysis and also categorizing tweets based on sentiments.[4] attempts to predict the revenue of movies by various techniques for estimating tweet frequencies .

Even though some effort has already been put in to predict the future with Twitter,in this project,we have tried to fine tune our model using other features like user interaction, data from other websites and user influence factor, apart from the number of user tweets to define popularity.

3. DATA

In this section, we will go through the various kinds of data sources used for this project.

3.1 Twitter

Our main source of data for this project was Twitter. To collect data for a movie, we searched for its official hashtag, instead of the movie title, to avoid ambiguities. The official hastags provided by twitter helped us to make sure that our data is cleaner and with less noise . We collected live tweets at regular intervals over a period of one and half months using the Twitter search API.

Our data collection was done in two ways. First, we collected data seven days prior to the release of the movie. This was done to predict the movie revenue on the weekend after the release. Our second set of data consisted of tweets five days prior to the movie release and five days after the movie release. Sentiment analysis on this data was done to show the reaction of users to the movie released, and to show whether the movie was liked by the users or not.

We collected data for 16 movies all of which have been released. This data for the movies were present as different file, one for each movie. To be able to do any kind of analytic work on our dataset, all these files had to be merged into one big file. This file was then fed as an input to our code to come up with various statistics.

Since Twitter data is highly unstructured and noisy, we had to clean and parse the data before deploying it in our project. There were several problems that we encountered while dealing with Twitter data:

- The tweet text was not all ASCII data. In addition to characters from different languages, there were emoticons as well in the data. A solution to this was to import this data in Latin 1 encoding, while reading in Pandas.
- The timestamp for the tweets collected did not have the same format. Some tweets had 'dd/mm/yyyy' for-

Movie
By The Sea
James White
Love the Coopers
My All American
Spectre
The 33
Secret In Their Eyes
Mockingjay 2
The Night Before
Carol
The Peanuts Movie
The Hallow
Creed
The Good Dinosaur
The Danish Girl
Victor Frankenstein

Table 1: List of movies for data collection

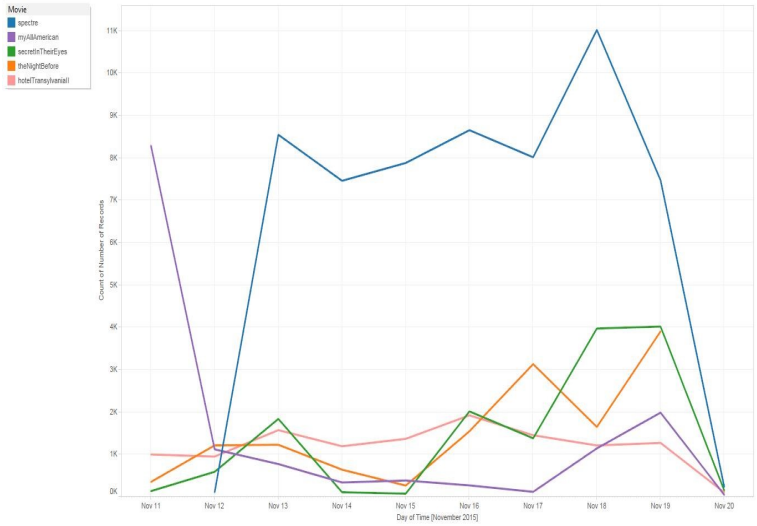

Figure 1: Time series of tweets for movies

mat, while others had 'mm/dd/yyyy' format. This required a lot of manual effort to make sure all date formats are the same.

Table 1 shows the movies for which the data has been collected and analyzed in this project.

3.2 BoxOfficeMojo

Data pertaining to the revenue of movies, the number of theaters in which a movie is running, as well as the weekly and weekend revenue statistics were collected from **www.boxofficemojo.com.**

4. METHODOLOGY AND RESULTS

Our initial goal was to predict box office revenues for a movie, based on the number of tweets made about that movie. However, predicting the complete box office revenue didn't seem to be a practical option for this project, because after the movie is released the revenue trend depends on a lot of other factors like critics' and users' reviews, post marketing strategies, etc. So, we restricted ourselves to predicting the box office revenue of a movie in the first weekend after it was released. We have taken samples where all the movies have been released on a Friday, and the box office revenue predictions are done till Sunday.

Firstly, we observe the number of tweets per day for different movies. Figure 1 shows the trend for the same. We have taken the top five movies which have the maximum number of tweets. As is clearly visible from the figure, the movie Spectre shows an astonishing trend in the number of tweets it received. This shows how enthusiastic people were about its release.

Predicting Movie Popularity:

Here, we will try to present a better approach to find out which movie is being talked about the most, or in other words, which movie is the most popular one currently. To do this, we have taken the following parameters into consideration:

- **Number of tweets received per movie:** This corresponds to what we discussed previously, with respect to Figure 1. According to the figure, we see that Spectre gets the maximum number of tweets over a span of eight days.
- **Number of users tweeting about a movie:** We plotted a graph for the number of users talking about a particular movie. Figure 2 shows a nearly power law plot of the same. We were expecting this plot to be a power law curve, but due to less data, the plot isn't perfect. We can see here as well, that Spectre is the movie that is being talked about the most, leading with a huge number.

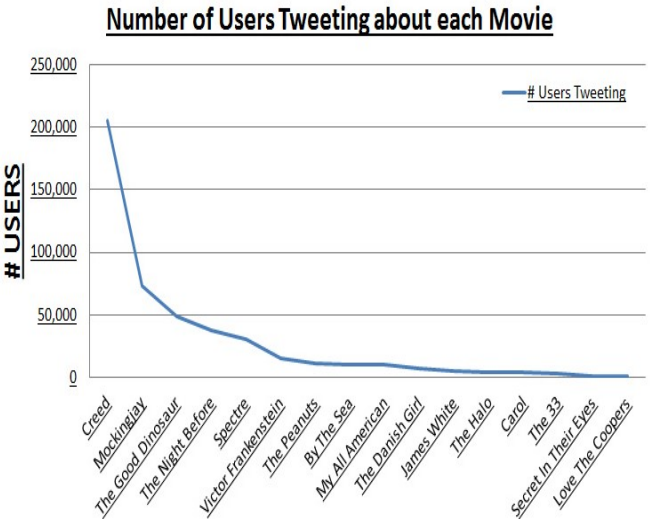

Figure 2: Number of users talking about a movie

- **Movie tweets over geographical location**

  We plotted our data pertaining to movies over geographical locations. The latitude and longitude data

were provided with the tweets. We do this to find out which movie is being watched most all over the world. Figure 3 shows this representation. No surprise this time, Spectre has wiped all grounds here, securing the first place yet again. Figure 3 shows the count of the unique users tweeting about a particular movie all around the world.

Inferences

Some points worth noting are:

- The maximum number of unique users are coming from California in USA for the movie "The Night Before". The other users are sparsely scattered in the eastern coast of USA. This movie has not been talked about a lot in the other parts of the world.
- As seen from Figure 1, we can see that the most tweeted movie is "Spectre". Although this is a big budget Hollywood movie, it has viewers all around the world, like the other James Bond movies. This can be seen by the blue dots all over the map. The people who tweeted about this movie don't largely belong to similar places, thus the dots are small, but come majorly from North America, Europe and South East Asia.
- On the other hand "Mockingjay 2" is being watched and tweeted about in North and South America and some parts of Europe. Some outliers can be observed in this case coming from Russian Territories and also from a single location marked by a medium sized red bubble.

From the analysis it can be seen that Spectre is the movie that is most talked about currently. Go Bond! Figure 4 shows a word cloud that shows the tweets' contents for Spectre. We can see that all sentiments shown here are positive, affirming our claim that Spectre is the most popular movie. Figure 5 shows how tweets can be used to predict the success of a movie in the coming weeks, the graph shows the percentage change of revenue and number of tweets from first to second weekend. Except for Love The Coopers we see a similar change between the two factors. It is clearly visible from the exploratory data analysis,how tweets are helping us to predict and gain some precise information about movies.

5. MOVIE REVENUE PREDICTION

Our data contains attribute values belonging to different ranges. To determine co-relation between such data poses a problem, in the sense that an attribute with a smaller value may not have as much a significant impact on the response variable as an attribute with a much larger value. For instance, the revenue for movies is typically dealt with in millions and billions. For this, the attribute average tweets per day can be only in tens of thousand, at the max, whereas the theater count in mostly in hundreds or thousands. This would skew our predictions. Hence, we standardized our data to give equal weight to all attributes. The standardization formula we used is as follows:

$$X' = \frac{X-\mu}{\sigma}$$

where X' is the standardized value, μ represents the mean of the values and σ represents the standard deviation with respect to the mean.

We saw a strong correlation between the average number of tweets per day(ATPD) and the box office gross with a p-value < 0.000205. There exists a strong linear dependency between these two terms- we were able to obtain a R^2 value of 0.811. This can be seen from the fact that Spectre which received the maximum ATPD of 19729 had a grand opening weekend with gross crossing $70.4M. On the other hand The Hallow received the minimum ATPD of 55 and hence had a low gross opening of $1559.

In order to strengthen the prediction model we define a new feature.

$$\text{Impact Factor}(IMF_m) = \sum_u (f)$$

IMF gives the information about how the unique users u who have tweeted about a movie m have an impact on the f followers of that user u. In simple terms the idea is if a user tweets about a movie we are assuming that all that who follow the user are impacted by what he tweets, could be positive or negative. With the introduction of this new variable, the R^2 value increases to 0.85. We further see that some movies like By The Sea has a very high IMF and ATPD, but still did not have a good opening,one reason could be because of the famous star cast of the movie but a bad plot to consume. We also see that the theater count for the movie By The Sea is only 10, this leads to another interesting fact that the box office revenue also depends on the theater count. So we add another feature variable for theater count(TC) into the prediction model and it is observed that the R^2 value increases to 0.87.

Figure 6 shows us a plot for the line that fits our sample data and our regression analysis.

However, taking into consideration only the impact factor isn't enough. This is because if a person has a negative comment, and has a large number of followers, it will only contribute more and more towards the prediction. But this is not correct. If there is a negative tweet, that negative factor should be discounted from the impact factor(IMF),i.e we need to take into consideration the sentiment of the tweet. We performed a sentiment analysis on the emotions of the tweet, that is, whether it is a positive tweet, or a negative tweet, or a neutral tweet. We will see ahead how we incorporated these two factors in revenue prediction.

6. SENTIMENT ANALYSIS

We felt that the count of the tweets were just not enough to correctly predict the revenue. We needed some sort of tool to gauge the opinions or feelings conveyed by these tweets. Thus we decided to use a naive sentiment analysis approach.

6.1 Preprocessing:

- The preprocessing of the data was done in R wherein data was cleansed and prepared for further processing.
- The non-ASCII characters were removed and only tweets with english words were extracted.
- Various Control characters and punctuation were removed too. We also removed links specified in the tweets.

6.2 Methodology:

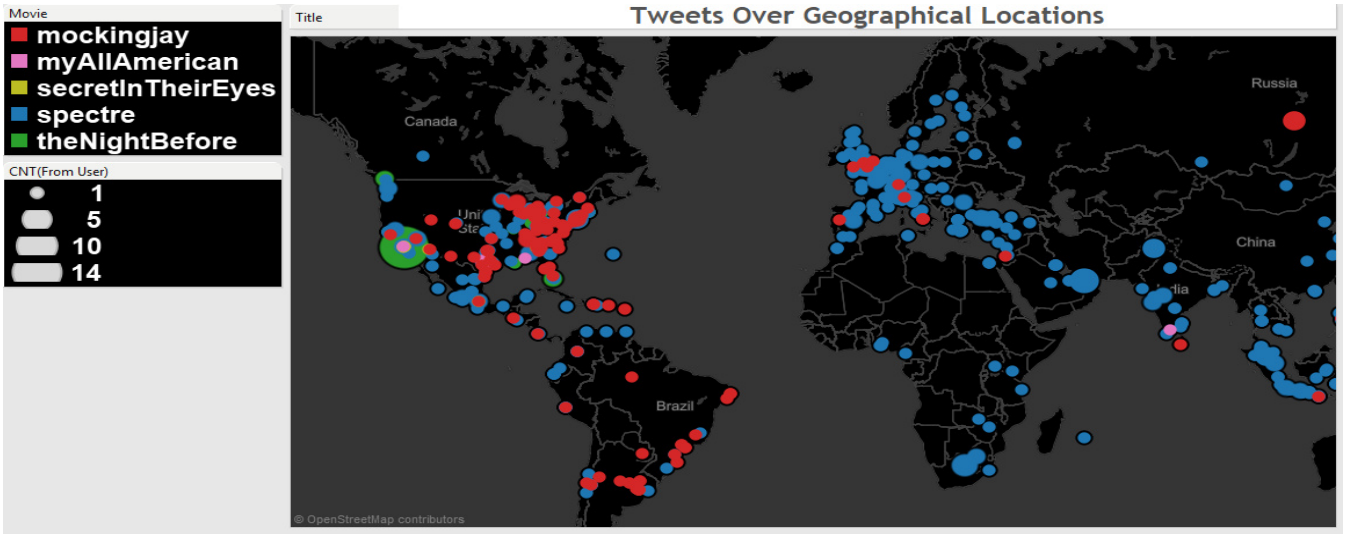

Figure 3: Tweets over geo-locations

Figure 4: Word cloud for Spectre

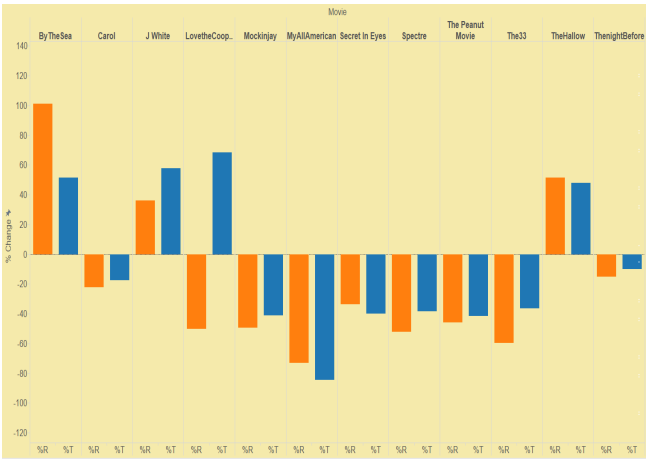

Figure 5: Tweets vs Revenue

- We utilized two sets of pre-labeled positive and negative words to compare the words present in the individual tweets.
- A score was assigned to each entry based on the number of positive or negative words present. eg. if two positive words were present and a negative word was present, the score would be +1.
- Now a cumulative score was calculated using the formula :
  Score= $\frac{\sum PositiveScore}{\sum PositiveScore + |\sum NegativeScore|}$
- The **Score** is used to find out the percentage of positivity of the tweet and thus the polarity of the sentiment conveyed.

6.3 Drawbacks and Assumptions:

There are a few assumptions and drawbacks of using this naive algorithm to compute the polarity of the tweets as shown below.

- **Ambiguous sentiment words** - "This movie is terrible" vs. "This movie is terribly good". the example above shows that there are two ways to say a particular statement and thus the algorithm [which does word comparisons from a pre-labelled set] would just search for a positive or negative word without taking it's semantics into consideration.
- **Missed negations** - "I would never in a millions years say that this movie is good and worth watching". This example shows how negations (sarcasm) is not taken into consideration.
- **Quoted/Indirect text** - "The reviews were terrible, but I disagree!" This example has both negative and positive words, but it's score would be 0 i.e. neutral, even though in actuality it is a positive statement.

The sentiment Score (S) obtained from the above naive approach didn't help us much, as the score for movies with great reviews on social media had a poor score, which on manual checking were the effects of the above drawbacks

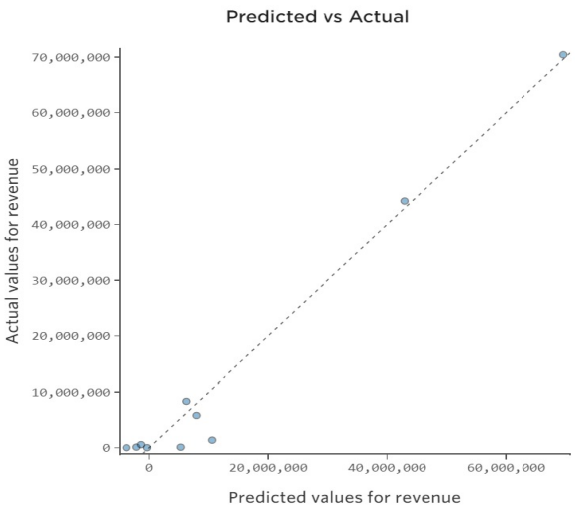

Figure 6: Regression Analysis Plot

limited.

We then adopted a different approach described in [5]. The paper looks into different methods to improve the Naive Bayes Classifier for sentiment analysis, features like n-gram, effective negation handling resulted in a significant improvement of accuracy. We were able to do a POST request on servers hosted by Vivekn [6] to get the confidence and sentiment of a tweet. A new score S' is calculated using the same formula used for S.

We define a new term at this point, Polarity:

$$\textbf{Polarity}= \frac{\sum PositiveSentiment}{\sum Sentiment}$$

Polarity is used to measure the success of a movie with respect to the sentiments in the tweets. We calculate the polarity of movies a week before the movie is released and a week after the release to see how are the audience accepting the movie. If polarity post release is > 1, then the audience seem to be liking the movie, if is $>> 1$, then even better, more people are taking positive about the movie and can hope it be a hit. Figure 7 shows the polarity of movies, pre and post release. Love The Coopers shows a polarity decrease from from [3.5] to [2.2], the revenue collected from boxofficemojo.com also show a decrease from [1.3M] to [.3M] in the second weekend.

As we can infer from above, the sentiment of a tweet might have some correlation affect on the prediction. We used this new sentiment score(S^*) and took a product of S^* and IMF to further check the R^2. Now, as is well know, adding more attributes to a regression model increases the R^2 every time, without taking into consideration whether or not this new attribute has any actual impact on the model or not. To check this impact, we checked the adjusted R^2 and predicted R^2 values. Table 2 shows a summary of the different R^2 values.

To be able to understand this table better, we shall describe each column separately. The first column show the feature that have been added to the model, in an incremental manner. R^2 value is the coefficient of determination. It states how close is the data to the fitted regression line. Adding new attributes to a regression model mostly increases the R^2 value, irrespective of whether or not the attribute has an impact on the model. To regularize this, R^2 adj value is used. Typically, this value should be less than R^2. If it is more, it means that the new attribute is improving our model by fluke, there is no actual relation. Lastly, R^2 pred indicates how well our model will predict values for new observations.

As is clear from the table, with the addition of every new attribute, our regression model is only improving. Hence, we were correct in considering a total of four attributes for our regression model. We have got a decent R^2 pred of 83%.

Feature	R^2	R^2 ***adj***	R^2 ***pred***
tweets	0.811	0.7921	0.75
tweets + theatre count	0.86	0.8577	0.8227
tweets + theatre count + S' * IMF	0.8829	0.862	0.8314

Table 2: Comparison of different R^2 values

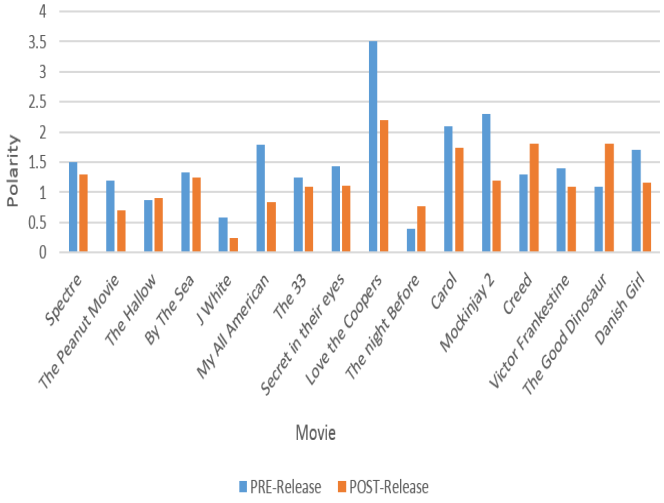

Figure 7: Polarity of tweets prior to and after the release date

7. EVALUATION

Building a regression is not sufficient to claim that our predictions are true. There has to be a way to evaluate our results. We took actual revenues from the official movie site BoxOfficeMojo and plotted them against the revenues predicted by our model. Figure 8 shows the plot for the same. The RMSE for our predicted values was 0.22.

8. CONCLUSIONS

For this project, we were limited by the number of movies for which the data could be collected. Due to the constraints imposed by Twitter on the number of days and the number of tweets to be collected,we could analyse only those movies which were released this month. This enabled us to have a minimum of 7 days worth of data for each movie . This

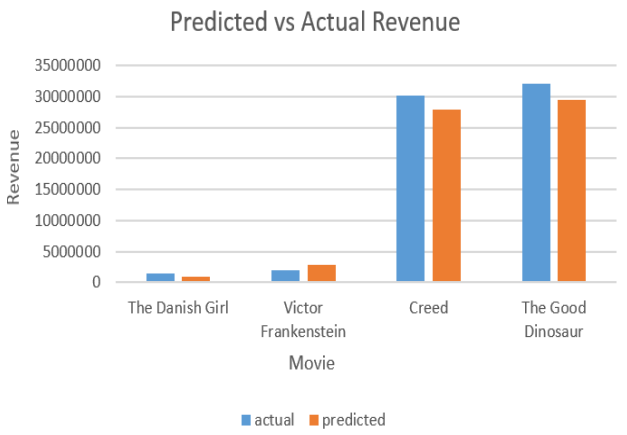

Figure 8: Predicted vs Actual Revenues

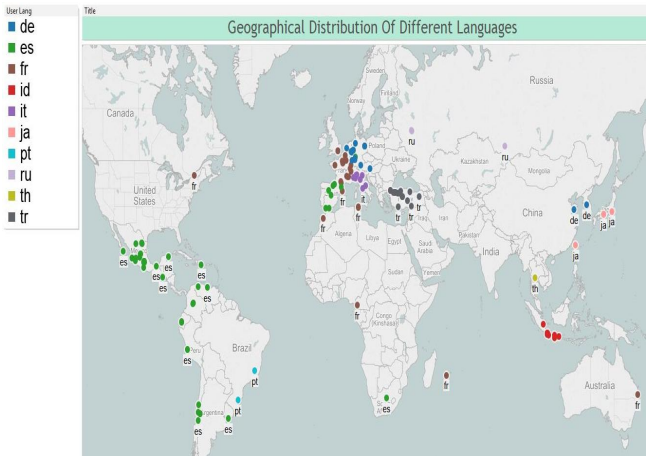

Figure 9: Language distribution for Spectre

is one of the major reason why our prediction was not very accurate. this small number of movies (16 movies) might have caused over-fitting of the curve. To reduce this effect we increased the number of features to better describe our data set. The few things we looked into are:

- **Tweets a week before the release** : Analysing this data set would give us a clear picture as to the popularity of the movie and the anticipation ensuing its release among the users.
- **Sentiment Analysis** : the count of the tweets is not enough to understand the movie's popularity. We need to understand the polarity of the tweets too. First we went ahead with a very naive method of sentiment analysis, wherein each word was compared to a list of positive and negative words and scored on the frequency of negative vs positive words occurring. This had its limitations and to overcome this we went ahead with a naive bayes method of n-gram sentiment analysis. This could compare phrases instead of individual words and some semantics could be inferred correctly, making the scores much more relevant.
- **Theatre Count :** We also found out that the above two things are not the only factors which would influence the movie revenue. If a movie is popular but has limited theatre release, the revenues would suffer. This could be seen for "James White" and "Carol".
- **Impact Factor :** To keep in mind the quality of the tweets , we added a weight to the sentiment score derived too.

So far in this project, we have seen how we can predict revenue of a movie based on certain factors such as number of tweets, number of theaters in which the movie was released, sentiment (polarity) of the tweets and the impact factor. We also concluded from our exploratory data analysis that Spectre is the "hottest" movie on Twitter as of last week, but now as the new movie releases come up, the twitter chatter about "spectre" has given way to "creed" and "The good dinosaur". This shows us how the trend of the movies show an upward curve in the beginning till its release and slowly die down and newer movies take its place.

There were some other simple things we looked into and would definitely be something we could work on in the future.

- We took the count of the tweets of the movie Spectre and plotted tweets of the top 10 non-English languages. The basic intuition was people who tweet in their regional languages would not have English as their primary language. So there might be a market to launch this movie dubbed in the native language or maybe with subtitles.
- As expected Mexico and parts of South America and Southern Europe have a lot of Spanish speakers and thus Spanish is shown to be popular language of discourse on Twitter.
- There were some surprising results also seen. Some French speakers are present in Africa and similarly some German speakers in the China Border. There are some trends in diversity which can be inferred from this.

With these attributes in place we could reach a good evaluation score and thereby build a good model.

9. REFERENCES

[1] Vasu Jain:*"Prediction of Movie Success using Sentiment Analysis of Tweets"*

[2] Andrei Oghina, Mathias Breuss, Manos Tsagkias and Maarten de Rijke. (2012) *Predicting IMDB movie ratings using social media.* Proceedings of the 34th European conference on Advances in Information Retrieval, pp. 503-507.

[3] Sitaram Asur and Bernardo A. Huberman, "*Predicting the Future with Social Media*",

[4] Devin Guillory and Chip Mandal "*Using Twitter Data to predict box office revenues.*"
[5] V. Narayanan, I. Arora, and A. Bhatia. 2013. *Fast and accurate sentiment classification using an enhanced Naive Bayes model.* In LNCS 8206, pp 194âĂŞ201
[6] https://market.mashape.com/vivekn/sentiment-3