**Term Project** **Text Mining for Social Media**

# Prediction of Movies’ Box Office Performance using Social Media

**Ojas Juneja**
**Saurabh Patel**
**Ronak Bhuptani**
**Date: 8th May, 2016**
**Instructor: Edmund Yu**

# Abstract

Social media content contains rich information about people's preferences. An example is that people often share their thoughts about movies using Twitter, Facebook, Instagram and YouTube. We are planning to do data analysis on tweets, Facebook posts and YouTube posts about movies to predict several aspects of the movie popularity before its release.

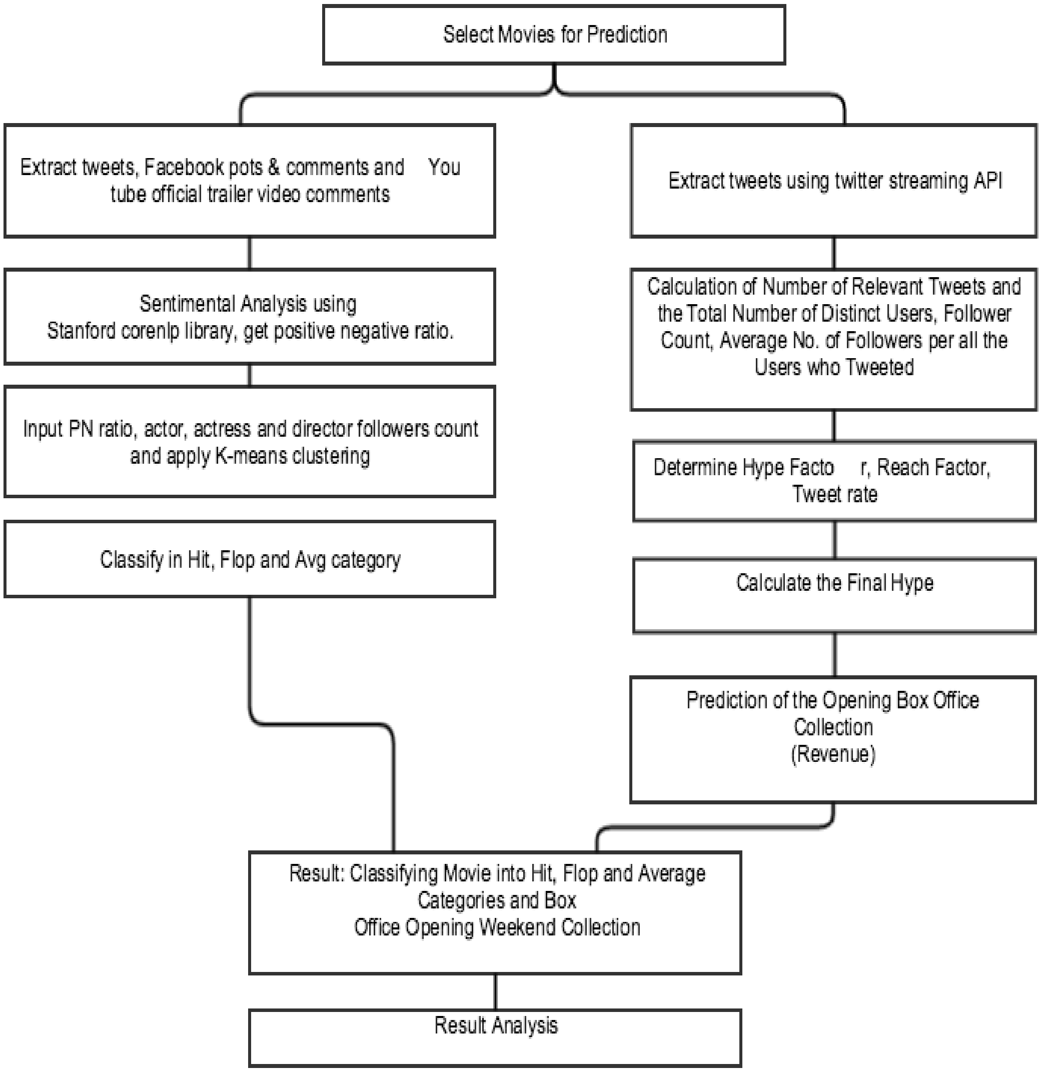
In this project, we are going to implement how social media content can be used to predict real-world outcomes. Box-Office performance of any movie can be determined by the amount of attention it gets before the movie is released. People's sentiment toward a particular matter when expressed online, can be very useful in many cases. The volume of discussion about products on Twitter can be correlated with the product's performance. It is also known that social network users represent the aggregate voice of millions of potential consumers. In particular, we will use tweets from Twitter to forecast box-office revenues for movies. For this, we have read several research papers and term papers which already contains information about this problem. After analyzing those papers, we came to know that Box-office performance of a movie is mainly determined by the amount the movie collects in the opening weekend. This amount is depending on following factors: Pre-Release hype is an important factor as far as estimating the openings of the movie are concerned. This can be estimated through user opinions expressed online through tweets, Facebook posts, YouTube comments etc. Each user is entitled to his own opinion which he expresses through his tweets. In addition to that user's comments, tweets sentiment analysis also help us to decide that movie is going to be success or not. Moreover, leading actor, actress and director follower's count on twitter and their rating also effects box office opening predication.

The project is divided in two principle parts. In the first part of the project, we are estimating box office opening weekend collection by its pre-release hype factor, tweet rate, total number of shows per day on all screens and average house full box office collection per screen. The second part is involved with sentimental analysis of tweets from Twitter, comments from Facebook page posts and YouTube official trailer videos. We also have considered popularity of movie actor, actress and director of respective movie.

# Methodology

The major work done in this project was to predict the box office movie performance and opening weekend box office collection of any movie. We had considered movies released in USA only for this project. We had considered movie-related tweets as an input. For Box office prediction, we have considered number of movie related tweets during one week before the movie release, Hype Factor of a movie, Tweet rate, number of screens and its houseful collection. This is our first part where we have predicated movies box office collection of first weekend using attention and hype factor. For second part of our project which is movie performance, we had used two components to classify movie in three different categories like hit, flop and average. First component was a set of movie tweets from Twitter web sites and did sentimental analysis on those tweets and second component corresponding to movie actor/actress/director rating.

The flow proposed system is shown in below figure. Proposed system initially selected few movies for prediction before release. The project work was broadly classified into following modules for application development.


Select Movies for Prediction
Extract tweets, Facebook pots & comments and You tube official trailer video comments
Sentimental Analysis using Stanford corenlp library, get positive negative ratio.
Input PN ratio, actor, actress and director followers count and apply K-means clustering
Classify in Hit, Flop and Avg category
Extract tweets using twitter streaming API
Calculation of Number of Relevant Tweets and the Total Number of Distinct Users, Follower Count, Average No. of Followers per all the Users who Tweeted
Determine Hype Facto r, Reach Factor, Tweet rate
Calculate the Final Hype
Prediction of the Opening Box Office Collection (Revenue)
Result: Classifying Movie into Hit, Flop and Average Categories and Box Office Opening Weekend Collection
Result Analysis

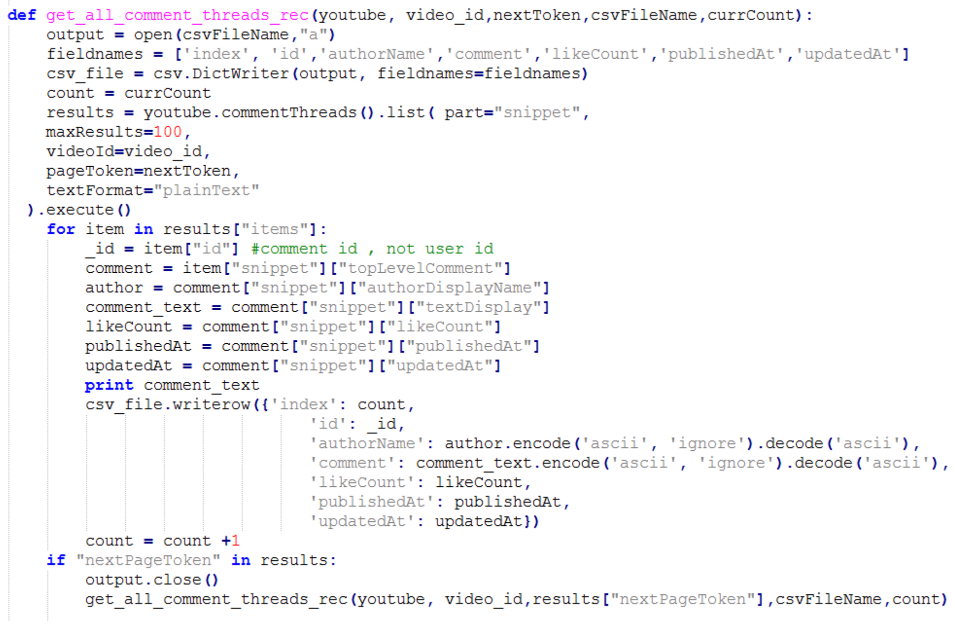
# Data Collection

### Extraction of data from Twitter

We have extracted only movie related tweets from the twitter before the release of the particular movie and we have extracted tweets for 12 movies. We have used Amazon EC2 instance to run all our tweets and we have created one script which will query over all the movie related keywords and based on the found keyword, it will stored in its respective file which is a csv file containing tweet text, user id, followers count of that user, time stamp.

**Extraction of data from YouTube**

We have also extracted comments from YouTube trailer videos in which we have used YouTube Data API v3. YouTube Data API allows only 100 comments in one result query. To overcome this drawback we have written python script such that we can request comments recursively by feeding next page token to the script until all the comments are fetched. Below is the snippet code of the same.



```
def get_all_comment_threads_rec(youtube, video_id,nextToken,csvFileName,currCount):
    output = open(csvFileName,"a")
    fieldnames = ['index', 'id','authorName','comment','likeCount','publishedAt','updatedAt']
    csv_file = csv.DictWriter(output, fieldnames=fieldnames)
    count = currCount
    results = youtube.commentThreads().list( part="snippet",
    maxResults=100,
    videoId=video_id,
    pageToken=nextToken,
    textFormat="plainText"
  ).execute()
    for item in results["items"]:
        _id = item["id"] #comment id , not user id
        comment = item["snippet"]["topLevelComment"]
        author = comment["snippet"]["authorDisplayName"]
        comment_text = comment["snippet"]["textDisplay"]
        likeCount = comment["snippet"]["likeCount"]
        publishedAt = comment["snippet"]["publishedAt"]
        updatedAt = comment["snippet"]["updatedAt"]
        print comment_text
        csv_file.writerow({'index': count,
                               'id': _id,
                               'authorName': author.encode('ascii', 'ignore').decode('ascii'),
                               'comment': comment_text.encode('ascii', 'ignore').decode('ascii'),
                               'likeCount': likeCount,
                               'publishedAt': publishedAt,
                               'updatedAt': updatedAt})
        count = count +1
    if "nextPageToken" in results:
        output.close()
        get_all_comment_threads_rec(youtube, video_id,results["nextPageToken"],csvFileName,count)
```

**Extraction of data from Facebook:**

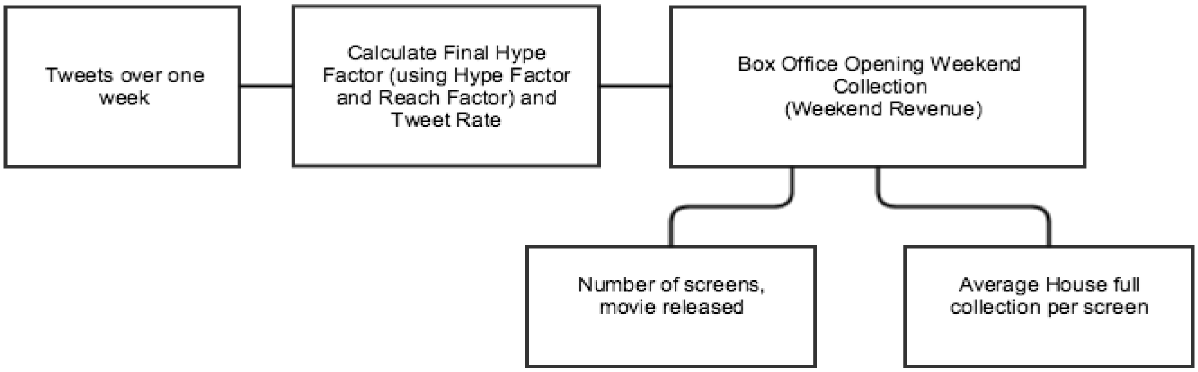
We have extracted posts, comments and likes for the particular page. Facebook don't allows multiple connection from the same IP Address. So, we extracted data one by one. Facebook JSON data involves pagination so to extract all the comments and posts, we have to go till last page and extract posts and comments.

Due to limitation of time, we have only considered twitter data in our system.

# Main Tasks of Project

## 1. Movie box office prediction using hype factor(opening weekend collection).

The prediction of movie box office opening weekend collection method was based on pre-release hype factor, tweet rate, total number of shows per day on all screens and the average price houseful collection of screen. Different users have their own opinion, which they express through simple tweets. We setup analyzes the opinion mining in these tweets with respect to a movie prior to its release in theaters. Estimate the hype of movie surrounding it and also predict the box office openings of the movie. Working of method used for implementation is shown in below figure and detailed steps are discussed below:


Tweets over one week
Calculate Final Hype Factor (using Hype Factor and Reach Factor) and Tweet Rate
Box Office Opening Weekend Collection (Weekend Revenue)
Number of screens, movie released
Average House full collection per screen

Working of method used for implementation is shown in above Figure and detailed steps are discussed below:

1. We have collected tweets of all movies of a week before it was released to calculate the hype factor. Below table shows how many tweets are collected for all movies in this time frame.

| Number | Movie Name | Relase Date | Total Tweets Collected |
|---|---|---|---|
| 1 | The Jungle Book | 15th April | **198623 (199K)** |
| 2 | Demolition | 8th April | **37219 (37 K)** |
| 3 | Fan | 15th April | **115297 (115K)** |
| 5 | Hardcore Henry | 8th April | **32500 (33K)** |
| 6 | Before I Wake | 8th April | **4145** |
| 7 | Louder Than Bombs | 8th April | **2099 (2K)** |
| 8 | Criminal | 15th April | **6774** |
| 9 | Green Room | 15th April | **14552** |
| 10 | A Hologram for the King | 22th April | **3064** |
| 11 | Elvis & Nixon | 22th April | **3028** |
| 12 | The HuntsMan | 22th April | **65872** |

2. For movie box office opening weekend collection initially we calculated final hype factor. For final hype factor need to find out movie hype factor, movie reach factor and tweet rate.

3. For initial movie hype factor need to find the total number of relevant tweets, number of distinct users those posted the tweets. The number of distinct users can be calculated by counting user-id of the users. This process starts one before the release of the any movie. The following formula is used for calculating the hype factor (α):

$$\alpha = \frac{Number\ of\ distinct\ users}{Number\ of\ tweets\ by\ all\ users}$$

The reason to use the ratio based approach is to get the closest approximation of hype which would be difficult had it been without the number of users. Number of distinct users is an important factor because hype can be best known through the number of users being interested in particular movie. The success of movie at the box office can be best determined by the ratio of number of users to the number of tweets rather than taking only number of tweets in consideration which would not give the best possible approximation. This is because the number of tweets is not directly proportional to the number of users. It is difficult to determine the accurate hype by only considering the number of tweets because the number of users is not known. The number of users is very important because ultimately the people will decide movie's success or failure. For example, Let the number of users who have posted the tweet be 10 and the number of tweets being posted by them pertaining to a movie be 100. By considering only, the tweets, the actual hype would not be determined. So, only after consideration of the number of user's actual hype is known. In this case since there are 10 users who have posted 100 tweets, after analyzing the number of users the fact that the hype is much less and movie is less likely to be a successful one is known. Another situation could be 500 tweets posted by only 50 users which prove that considering only tweets is specious.

4. To enhance the estimation of hype it is necessary to consider the reach of a particular tweet by including the follower count of a particular user who referred to the movie in his tweet if the count is above a certain thresh-hold value (τ).The follower-count factor can also be a considered a factor to ensure the reach factor is also included in the determination of the hype. The reach factor (σ) can be given as

$$\sigma = \frac{(followers\ count - \tau)}{followers\ counr}$$

Where τ = average no. of followers per all the users who tweeted. σ can be scaled down to a scale of 0.1-1 with 0.1 being assigned to the thresh hold value assuming cases where the follower count being more than 10 times the thresh hold value as a rare case and assigning it the value 1.

5. Hype factor formula, the hype factor gives values which may be an integer or decimal and we can see that Hype can be determined from the initial hype factor and reach factor.

$$\text{Hype} = \frac{\alpha + \sigma}{2}$$

6. In our project, we have calculated tweet rate over the one-week period for per hour. It helps to make correct predication about box office opening. We introduce this new factor which we have not found in relevant work papers. The number of tweets per hour is another factor which gives us a rough idea about the hype, the movie possesses as it would mean that there are a lot of tweets being posted about the movie which means that the movie is being discussed and anticipated for. Also the reach of the tweet is to be considered as it would mean that that a wide number of followers were subjected to witnessing a tweet posted by the person they follow.

With consideration of these factors the hype is calculated which should be as high as possible to ensure the film a good opening weekend.

$$\text{Tweet Rate} = \frac{Total\ Number\ of\ Tweets}{Number\ of\ Hours}$$

For few movies, we are getting very low tweet rate. We used threshold value for this and assigning 0.1 value where tweet rate is going too low.

7. Final Hype factor formula:

$$\text{Hype} = \text{Tweet Rate}* \frac{\alpha + \sigma}{2}$$

After collecting all tweets and calculating all the number we have come up with this inputs to calculate the box office prediction using below steps.

| Number | Movie Name | Relase Date | Hype (With) | Hype (Without) | Total Tweets Collected | Tweet Rate |
|---|---|---|---|---|---|---|
| 1 | The Jungle Book | 15th April | 0.55 | 0.734950967 | 198623 | 1 |
| 2 | Demolition | 8th April | 0.68950287 | 0.739260809 | 37219 | 0.178705909 |
| 3 | Fan | 15th April | 0.517959479 | 0.552847941 | 115297 | 0.576000896 |
| 5 | Hardcore Henry | 8th April | 0.731585997 | 0.731585997 | 32500 | 0.15469276 |
| 6 | Before I Wake | 8th April | 0.733150272 | 0.733150272 | 4145 | 0.1 |
| 7 | Louder Than Bombs | 8th April | 0.664871993 | 0.732716074 | 2099 | 0.1 |
| 8 | Criminal | 15th April | 0.685320827 | 0.735292138 | 6774 | 0.1 |
| 9 | Green Room | 15th April | 0.729369718 | 0.778498248 | 14552 | 0.1 |
| 10 | A Hologram for the King | 22th April | 0.599436023 | 0.711817603 | 3064 | 0.1 |
| 11 | Elvis & Nixon | 22th April | 0.574239612 | 0.729572047 | 3028 | 0.1 |
| 12 | The HuntsMan | 22th April | 0.608685653 | 0.681982878 | 65872 | 0.324504895 |

We have considered amount in $ for all movies except movie Fan which was released in India and for that we have considered amount in INR. The opening weekend collection can be calculated using the hype factor and the knowledge of how many screens the movie is going to release in the occupancy of each movie theatre is analogous with the hype surrounding the movie.

The opening box office collection (O) can be predicted as

$$O = \mu * Hype * \varphi$$

Where

O is the opening box-office collection.
µ is the number of shows per day in all screens together for the weekend.
$\varphi$ is the average price of all tickets per screen per show.

## Evaluation

We also calculated MSE of different movies. The different results are compared with the help of actual box office opening of first weekend from **BoxOfficeMojo** website. Here the employed errors are the MSE of the input data set. If Yt is the actual observation for time period t and Xt is the forecast for the same period. In below given table, predictor values are computed by using hype factor and other input values. The formula used for calculation of MSE is as shown in below MSE equation. Where N is number of movies, Yt is the actual observation and Xt is the forecast, in this case we have taken N as number of movies.

$$MSE = \frac{1}{N}\sum (Y_t - X_t)^2$$

Below is the predicated values of opening weeking box office.

| Number | Movie Name | Predicted Opening (with) in million | Predicted Opening (without) in million | Opening (Boxoffice) in million | Square (with) | Square (without) |
|---|---|---|---|---|---|---|
| 1 | The Jungle Book | 78.0945 | 104.355687 | 103.261464 | 633.376077 | 1.197323974 |
| 2 | Demolition | 3.709485 | 3.977179 | 1.100042 | 6.80919277 | 8.277917317 |
| 3 | Fan | 483.3191 | 515.874272 | 523.5 | 1614.504725 | 58.15172753 |
| 5 | Hardcore Henry | 12.027254 | 12.027254 | 5.107604 | 47.88155612 | 47.88155612 |
| 8 | Criminal | 6.481421 | 6.954025 | 5.767278 | 0.510000224 | 1.408368442 |
| 10 | A Hologram for the King | 0.847003 | 1.005798 | 1.138605 | 0.085031726 | 0.017637699 |
| 11 | Elvis & Nixon | 0.770916 | 0.97945 | 0.466447 | 0.092701372 | 0.263172078 |
| 12 | The HuntsMan | 26.395782 | 29.574331 | 19.445035 | 48.31288386 | 102.6026375 |

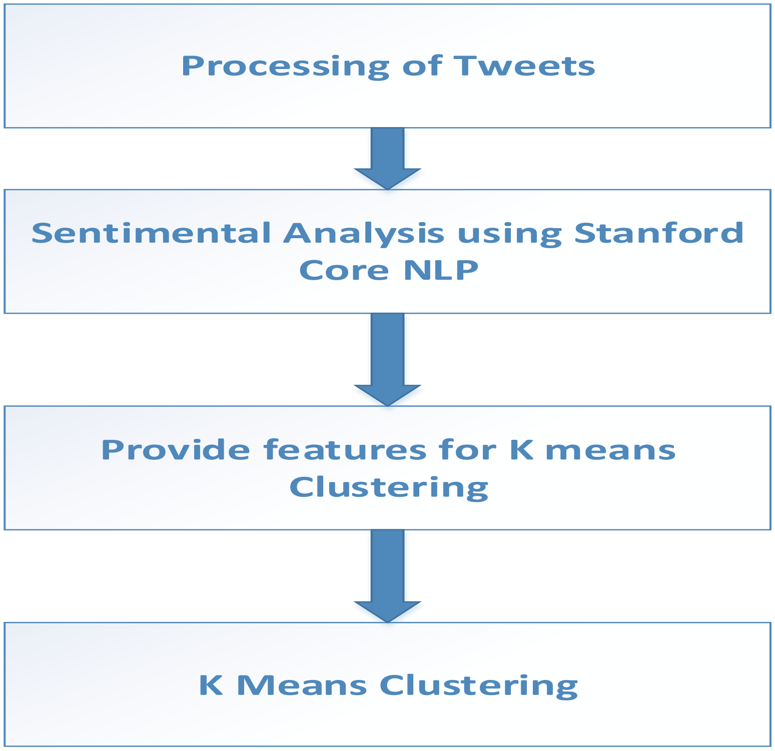
Below is the table for MSE,

| | | |
|---|---|---|
| **MSE (with)** | | **261.2857964** |
| **MSE (without)** | | **24.42226007** |

We finalized the model of without considering empty hours' hype, which gives us better prediction.

## 2. Movie Performance using Sentiment Analysis and popularity factors

Sentiment analysis refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. We have done sentimental analysis on twitter using Stanford Core NLP library which classifies data into positive, negative and neutral tweets. After that we have took sentiment factor and popularity of actor, actress and director to cluster the movies into 3 classes: hit, neutral and average. Working of this method is shown in below diagram:


Processing of Tweets
Sentimental Analysis using Stanford Core NLP
Provide features for K means Clustering
K Means Clustering

1. **Preprocessing of Tweets**
   The tweets we have obtained from data collection steps are processed because every tweet has some noisy words. It is necessary to remove the noisy words from those tweets. We have done the following steps:
   a) Remove English stop words.

b) Remove punctuation, special characters, user mentions and user id from tweets.
c) Remove URL, hashtags, replace movie name with “MOVIE” word as it gives problem for sentimental analysis. For example, Criminal, Die, Demolition words creates negative sentimental.
d) Used Porter Stemmer to remove morphological endings from words in English.

**2. Sentimental Analysis**

We have developed twitter sentimental analyzer using variety of methods. Since, we have Setimental140 training data which contains positive and negative tweets. Our models are able to classify tweets into positive and negative. Those models shows fair accuracy on training data and since we have unsupervised test data. So, we have manually labelled some of test data in order to measure accuracy on our test data and since the test data also contain noisy tweets. So, we are not able to achieve good accuracy on that particular set of test data.

Example of Noisy Data: **“**chance win x tickets movie London premiere”

Below table shows accuracy obtained using train Data.

| **Classifier** | **Input** | **F-Score** |
|---|---|---|
| SVM – linear kernel | Sentimental140–25K positive, 25K negative | 0.701 |
| SVM – RBF kernel | Sentimental140–25K positive, 25K negative | 0.72 |
| Naïve Bayes | Sentimental140–25K positive, 25K negative | 0.61 |

Stanford Core NLP removes all this drawbacks by classifying data into positive, negative and neutral and we do not require any train data for this.

**3. Generation of features for K means Clustering**

After doing Sentimental analysis on data we classified the movie into Hit, Flop and average by considering the factors:

a. PN ratio,
b. Actor popularity,
c. Actress popularity,
d. Director Popularity
e. PN ratio = Number of positive tweets &comments/ negative tweets & comments.
f. Director, Actor, Actress popularity = Number of followers on twitter.

Problems with collecting popularity data:

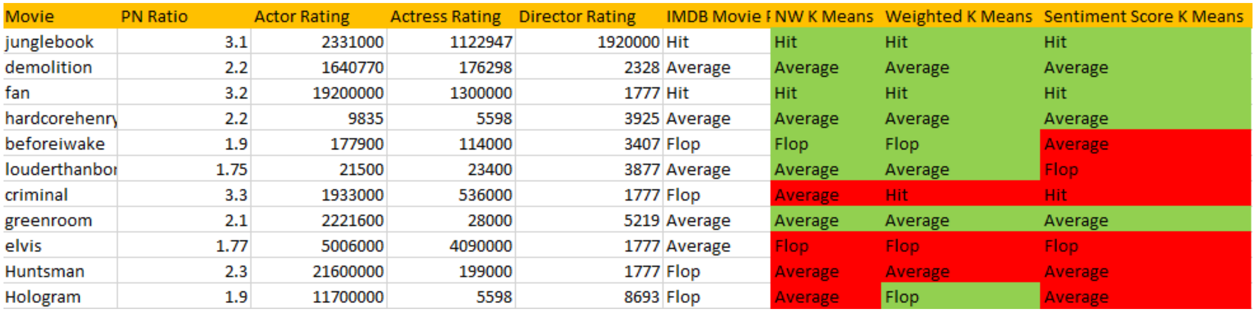
1) For some movies, there is no Facebook page or tweet page for director/Actor/Actress.
2) For some movies, there are more than one lead actors/actress.

Solution:

1) For those Actor/Actress/Director, we have considered the lowest followers in our dataset as the followers count of those Actor/Actress/Director count
2) For movies having more than one lead actor/actress, we have done the summation of all the lead actors, actress.

## 4. Prediction using K Means Clustering

We have followed different approaches to classify movies into HIT, FLOP and AVERAGE. Below table shows the clustering of movies and their PN Ratio and Actor, Actress and Director popularity.



| Movie | PN Ratio | Actor Rating | Actress Rating | Director Rating | IMDB Movie I | NW K Means | Weighted K Means | Sentiment Score K Means |
|---|---|---|---|---|---|---|---|---|
| junglebook | 3.1 | 2331000 | 1122947 | 1920000 | Hit | Hit | Hit | Hit |
| demolition | 2.2 | 1640770 | 176298 | 2328 | Average | Average | Average | Average |
| fan | 3.2 | 19200000 | 1300000 | 1777 | Hit | Hit | Hit | Hit |
| hardcorehenry | 2.2 | 9835 | 5598 | 3925 | Average | Average | Average | Average |
| beforeiwake | 1.9 | 177900 | 114000 | 3407 | Flop | Flop | Flop | Average |
| louderthanbor | 1.75 | 21500 | 23400 | 3877 | Average | Average | Average | Flop |
| criminal | 3.3 | 1933000 | 536000 | 1777 | Flop | Average | Hit | Hit |
| greenroom | 2.1 | 2221600 | 28000 | 5219 | Average | Average | Average | Average |
| elvis | 1.77 | 5006000 | 4090000 | 1777 | Average | Flop | Flop | Flop |
| Huntsman | 2.3 | 21600000 | 199000 | 1777 | Flop | Average | Average | Average |
| Hologram | 1.9 | 11700000 | 5598 | 8693 | Flop | Average | Flop | Average |

Below table shows clustering of movies and their accuracy using different approaches:

| **Approach** | **Input** | **Accuracy** |
|---|---|---|
| Non Weighted K Means | PNratio,Actor,Actress,Director popularity scaled to 0 - 10 | 64% |
| Weighted K means | PNratio scaled to 0 -20 and other attributes scaled to 0-10 | 54% |
| Weighted K means | PNratio scaled to 0 -30 and other attributes scaled to 0-10 | 62% |
| Weighted K means | PNratio scaled to 0 -40 and other attributes scaled to 0-10 | 72% |
| Sentiment Score K means | Only PN ratio is considered | 45% |

## Conclusion

1. To predict the box office movie collection, tweet rate and hype are very crucial factors.
2. To predict whether a movie is hit, flop or neutral, sentiment factor is important but it cannot predict alone. We will need other factors like popularity count to predict.

## References

1. https://www.scipy.org/
2. http://pandas.pydata.org/
3. Box-Office Opening Prediction of Movies based on Hype Analysis through Data Mining [ARTICLE in INTERNATIONAL JOURNAL OF COMPUTER APPLICATIONS, OCTOBER 2012]
4. Using Twitter data to predict the performance of Bollywood movies [Dipak Damodar Gaikar and Bijith Marakarkandy, Department of Information Technology, Thakur College of Engineering and Technology, Mumbai, India]
5. Predicting the Future with Social Media- S Asur, B Huberman, HP Labs, HP Journal, Jan 2012
6. https://developers.google.com/youtube/v3/docs/commentThreads/list
7. http://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/pipeline/StanfordCoreNLP.html

8. http://thinknook.com/twitter-sentiment-analysis-training-corpus-dataset-2012-09-22/
9. http://www.imdb.com/
10. http://www.boxofficemojo.com/
11. https://en.wikipedia.org/wiki/The_Jungle_Book_(2016_film)#Box_office
12. Presentation by Ojas Juneja, Ronak Bhuptani and Saurabh patel.