Journal of Advanced Computing and Communication Technologies ***(ISSN:* 2347 - 2804*)***
Volume No 5 Issue No. 1. February 2017

MOVIE RATING PREDICTION BASED ON TWITTER SENTIMENT ANALYSIS

By

[1]Mr.Abhishek Kesharwani, [2]Mr. Rakesh Bharti
M.Tech Scholar, UIT, Dr. A. P. J. Abdul Kalam Technical University, Uttar Pradesh, Lucknow, India
Asst Prof., UIT, Dr. A. P. J. Abdul Kalam Technical University, Uttar Pradesh, Lucknow, India
abhishekkesharwani689@gmail.com, goswami.rakesh@gmail.com

ABSTRACT
With microblogging platforms such as Twitter generating huge amounts of textual data every day, the possibilities of knowledge discovery through Twitter data becomes increasingly relevant. Similar to the public voting mechanism on websites such as the Internet Movie Database (IMDb) that aggregates movies ratings, Twitter content contains reflections of public opinion about movies. This study aims to explore the use of Twitter content as textual data for predicting the movie rating. In this study, we extract number of tweets and compiled to predict the rating scores of newly released movies. Predictions were done with the algorithms, exploring the tweet polarity. In addition, this study explores the use of several different kinds of tweet classification Algorithm and movie rating algorithm. Results show that movie rating developed by our application is compared to IMDB and Rotten Tomatoes.

General Terms
Movie rating prediction, tweet polarity

Keywords
Twitter sentiment analysis, comparison of twitter rating from IMDB and Rotten Tomatoes, tweet classification algorithm, movie rating algorithm.

1. INTRODUCTION
The task of movie rating prediction by extracting the tweets and hashtags related to particular movie. Twitter data can be accessed through the public API provided by the Twitter. These APIs can be accessed only by authentication requests, which must be signed with valid login ID and password. Twitter provides authentication keys for extractions of the tweets. Tweet extracted from twitter API is collected in MySql where the tweet is having a unique tweet id, twitter id and tweet date. Movie name is added by the user and the release date of the movie is also added by the user. We have to add the twitter Id of the movie from the twitter and all the tweets are extracted from the twitter. We have to update the tweets of the movie by establishing the connection with twitter from our application, and then all the recent tweets are added to our database. We have also calculated the popularity of the movie at the time we extract the tweets. Each time we want to predict the rating or calculate the popularity of movie labelling of every tweet is also calculated. Apart from the tweets obtained from Twitter, the application also calculates the sentiment associated with the tweet. The application also stores the sentiment associated with the tweet in the database. Sentiment of a tweet is categorized as positive, negative, neutral and irrelevant. In this application we developed a module that is used to create bag a words from the tweets of old movies. We create two bags of words positive and negative .This is the backbone of our application. Each tweet that we extract from the twitter should be categorised as positive or negative tweet. We have a module update positive and negative words in my application which is used for updating positive and negative bag of words. For calculating the movie rating we ignored neutral and irrelevant tweets as these are not useful for any type of information for movie review. For movie rating prediction we have designed a module by which you can select any movie and all the tweets related to that movie is loaded in the algorithm with hashtags of the movie is also added so that all the review is added for the particular movie and more precise and accurate rating is calculated. When we search for a particular movie which is twitter id of the movie then rating out of 10 is calculated and change to format of 5 star rating with the help of our algorithm we design for movie rating. The performance of this application is evaluated by comparing its results with results from popular movie rating websites like IMDB and Rotten Tomatoes. The rating calculated for various movies released by fetching tweets related to the movie is stored in our application and real time rating of the movie is calculated and is represented in tabular form and graphically. The ratings collected show that the Twitter rating application is following similar trends as shown in IMDB and Rotten Tomatoes.

Ratings from Twitter application are observed to be of a lower value. This can be attributed to the fact that the number of users rating a movie on the other websites is almost 100x times than what the Twitter application is using. A small experiment was conducted with varying number of tweets used by Twitter application to confirm this theory.

2. RELATED WORK
The topic of using social media to predict the future becomes very popular in recent years. Different work has been already done using twitter content for predicting the sentiment of tweets. Movie sentiment analysis based on public tweets [1] ,In this paper they introduced a special approach to understanding the tongue and frequency of words from one

sentiment category enabling a much better sentiment classification compared to the normal machine learning techniques. They additionally introduce an added sentiment category - the neutral category. In their analysis they use the Python programming language with the NLTK library and compare so obtained results with the normal machine learning. Predicting Ratings for New Movie Releases from Twitter Content [5] that is textual knowledge from Twitter will be seen as an in depth supply of data relating to a particularly broad form of subjects. With a lot of users actively expressing themselves on-line, a large quantity of information is generated daily. Since this knowledge for an oversized half consists of human expressions, Twitter knowledge will be seen as a valuable assortment of human opinion or sentiment, which might be mechanically extracted with comparatively high accuracy.

3. METHODOLOGY

The overall system can be designed in following phases

i. Tweet Collection
ii. Tweet Classification
iii. Rating Movies.

3.1 Tweet Collection

Twitter data can be accessed through the public API provided by the Twitter. These APIs can be accessed only by authentication requests, which must be signed with valid login ID and password. Twitter provides authentication keys for extractions of the tweets. We have to follow some steps to create Authentication keys.

i. Create application on twitter.
ii. Manage Application
iii. Change the permissions to read and write.
iv. Retrieve Authentication keys.

First we create application on twitter by signing in to https://apps.twitter.com/app/new. Second Step we have to manage Application. In the third step change the application permissions to read and write. After the completion of the Application ,we have the following unique keys that is required to fetch tweets from twitter

a. Consumer key
b. Consumer Secret key
c. Access token
d. Access token secret

Tweet Extracted from twitter having complete information like date of tweet, tweet ID, user ID, retweet count etc.We will use only tweet date, tweet ID and tweet. We will add two more columns release date and added by to have some extra information about the analysis of movie. We have integrated twitter API into our application so as to fetch all the tweets related to a particular movie and all the news and comments related to a particular movie. There are so many limitation related to this API as this will extract limited tweets approximately hundred tweets in one time. But the tweets are collected into the database as we have MySql database in our back end. We have to update the tweets of the movie by establishing the connection with twitter from our application, and then all the recent tweets are added to our database.

3.2 Tweet Classification

In this module I label every tweet as positive, negative, neutral and irrelevant. Each time we want to predict the rating or calculate the popularity of movie and labeling of every tweet is also calculated. Apart from the tweets obtained from Twitter, the application also calculates the sentiment associated with the tweet. In this we tokenize the tweet into different tokens separated by space and compare each token with our predefined set of positive and negative bag of words. After the comparison of tokens we find the total number of positive and negative tokens in the tweet. Count the total number of positive and negative tokens in the tweet and label them as p and n respectively. Calculate the value of ratio as total number of positive tokens to the total number of positive and negative token.

Table 1 Table for tweet labeling

Ratio	Tweet Label
ratio>0.5	Positive
ratio=0.5	Neutral
ratio<0.5	Negative
ratio=0/0 (p==0 && n==0)	Irrelevant

3.2.1 Bag of words

In this application we developed a module that is used to create bag a words from the tweets of old movies. We create two bags of words positive and negative .These are the backbone of our application. Each tweet that we extract from the twitter should be categorized as positive or negative tweet. We have a module update positive and negative words in my application. In this module if we want to add any positive or negative word ,if that word already exist in the database of bag of words then it is not updated.

3.3 Algorithm for Tweet Classification

1. Extract the tweet from twitter using Twitter API.
2. Store the tweet in the database. Each tweet has a unique id and date of tweet which is also store in the database.
3. For each tweet we have to specify the tweet label. Tweet labels are positive, negative, neutral and irrelevant.
4. Calculate the value of p and n i.e. total number of positive and negative words obtained by comparing each word stored in our manually collected positive and negative collection of bag of word.
5. Classify the tweet label of each tweet using the following formula

```
ratio= p/ (p + n)
If (ratio>0.5)
{
        Tweet label="positive";
}
If (ratio==0.5)
{
        Tweet label="neutral";
}
If (ratio<0.5)
{
```

```
        Tweet label="negative";
    }
    If (p==0 and n==0)
    {
        Tweet label="irrelevant";
    }
```

6. Here p and n are total number of positive and negative words obtained by comparing each word stored in our manually collected positive and negative collection of word set.

3.4 Rating Movie

Rating is calculated by the total number of positive tweets over total tweets. This was designed that the rating is as correct because it will get the most recent views about the particular movie and twitter chat about the popularity and criticism of the movie

3.5 Algorithm for Rating Movie

1. Update the tweets and hashtag of a movie with twitter id and store it into database.
2. Select all the tweets and hashtags from the database with same twitter id and apply tweet classification algorithm.
3. Classify all the tweets and store the tweets with tweet labels as positive, negative, neutral and irrelevant.
4. Ignore neutral and irrelevant tweets. Neutral tweet does not specify any positive or negative sentiment as they create a situation of ambiguity and irrelevant tweets do not specify any sentiment.
5. Calculate the value of rating as

   rating = ((total positive tweets) / (total positive tweets +total negative tweets))*10)

6.
```
if (rating<2.0 and rating >=0)
{
Movie rating = 1 star
}
if ( rating>=2 and rating<3 )
{
Movie rating = 2 star
}
if (rating>=3 and rating<4 )
{
Movie rating = 2.5 star
}
if (rating>=4 and rating<5 )
{
Movie rating = 3 star
}
If (rating>=5 and rating<6)
{
Movie rating = 3.5 star
}
If (rating>=6 and rating<8 )
{
Movie rating = 4 star
}
If (rating>=8 and rating<9)
{
Movie rating = 4.5 star
}
If (rating>=9 and rating <=10)
{
Movie rating = 5 star
}
```

4. EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

The performance of this application is evaluated by comparing its results with results from popular movie rating websites like IMDB and Rotten Tomatoes. IMDB is internet information of data associated with films, television, programs etc. the location permits registered users to rate any film on a scale of one to ten, except for writing reviews regarding it. The location displays a weighted mean of user ratings and displays it next to the movie title. This website has 6.05 Daily Page views per visitor [7] and a median of fifteen million individuals visit the web site per month. Rotten Tomatoes may be a web site dedicated to film reviews and news. It offers 2 sorts of scores for movies – Tomato meter critic mixture score and Audience score. The Critic mixture score reflects reviews and ratings from varied newspaper writers or from those that belong to film critic associations. The Audience score is calculated supported user's reviews and ratings. For the aim of confirmatory this application, solely Audience score is taken into account. Registration is free however the location requests permission to look at user's social network profiles. It gets 3.60 Daily Page views per visitor and a median of thirteen million individuals visit the web site per month. [8].The rating calculated for various movies released by fetching tweets related to the movie is stored in our application and real time rating of the movie is calculated and is represented in tabular form and graphically as shown in table 2 and figure 1.

Table 2 Twitter Rating vs. IMDB and Rotten tomatoes

Movie Name	Twitter Rating	IMDB	Rotten Tomatoes
Ae dil hai mushkil	2	3	2.9
Shivaay	3.5	3.5	3.8
Force2	3	3	3.2
Pink	3	4	4.2
Raees	2	4	3.9

Kaabil	4	4	4
Dangal	3.5	4.5	4.6
Kahaani2	4	3.5	3.7
Bahubali	2.5	4	4.3

As shown in the Figure 1 twitter rating for the movie "Ae Dil hai Mushkil" is 2 while that IMDB is 3 and rotten tomatoes rated it 2.9.For the movie "Kaabil" we predict the same rating as predicted by the IMDB and Rotten Tomatoes. For the movie "kahaani2" we predict rating as 4 but IMDB rated as 3.5 and rotten tomatoes rated 3.7.

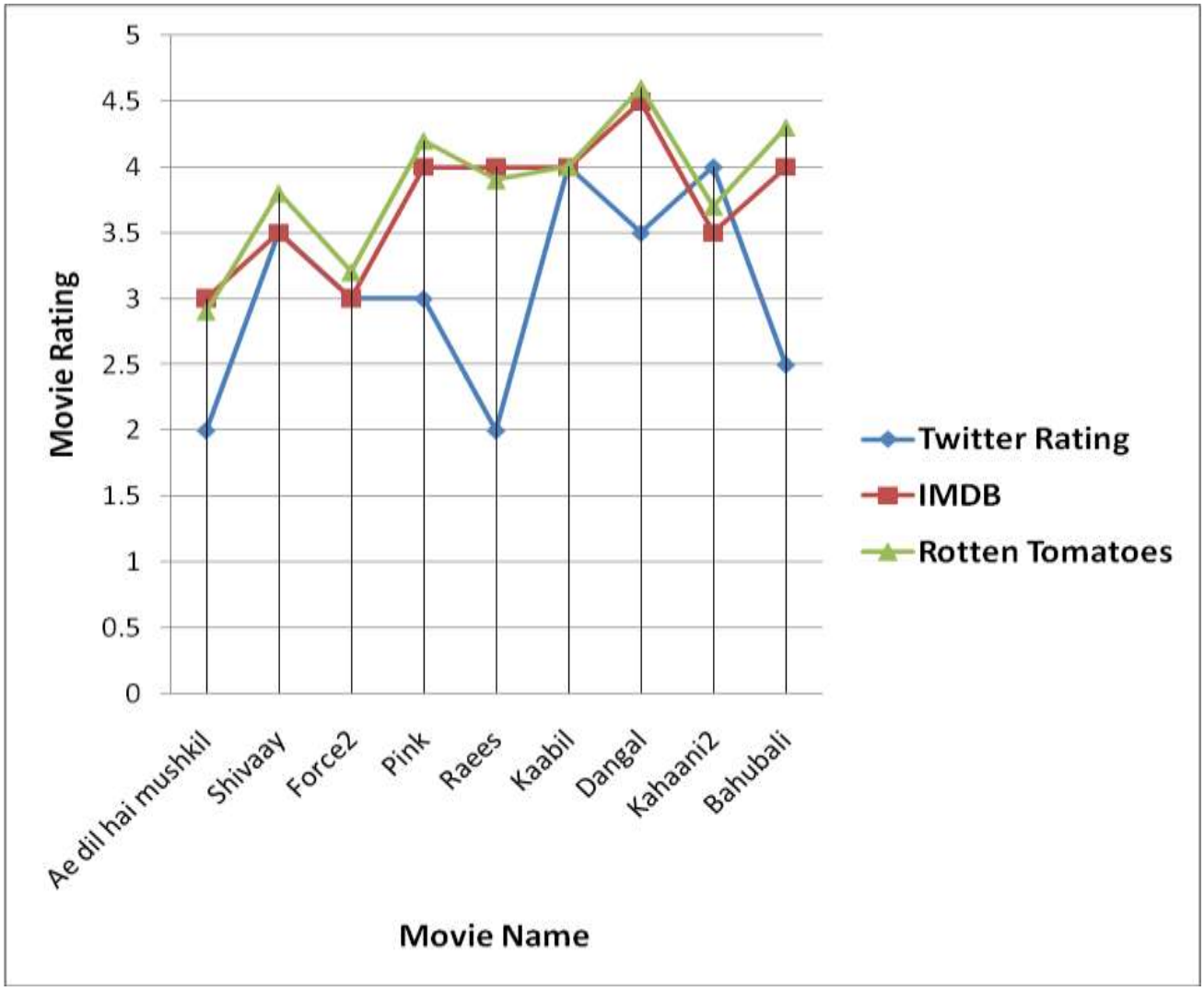

Fig 1: Twitter Rating versus IMDB and Rotten Tomatoes

Table 3 Twitter Reviews versus IMDB and Rotten Tomatoes Reviews

Movie Name	**Twitter Reviews**	**IMDB Reviews**	**Rotten Tomatoes Reviews**
Ae dil hai mushkil	1375	12644	921
Shivaay	392	6341	154
Force2	233	2245	50
Pink	199	18102	302
Raees	431	14390	332
Kaabil	1200	7462	131
Dangal	193	30720	1433
Kahaani2	200	1719	204
Bahubali	431	67698	1743

As show in Table 3 of Twitter Reviews and IMDB and Rotten Tomatoes reviews, the number of reviews are very large for IMDB as compared to twitter reviews for the movie "Ae dil

hai mushkil" total number of tweets we collected are 1375 while that of IMDB is 12644 and that of Rotten Tomatoes we obtained the ratings as 2 for twitter, 3 for IMDB and 2.9 for Rotten Tomatoes. For the movie "Kaabil" total number of tweets we collected are 1200 while that of IMDB is 7462 and that of Rotten Tomatoes 131 we obtained the ratings as 4 for all are same. The result shows that if we obtain huge amount of data from twitter than we calculate more accurate rating by our algorithm as this application works more accurately with huge amount of data, but currently the application can retrieve close to 100 tweets per API call, depending on how recently the movie was tweeted about. It has to be stressed that this is not enough data to determine the opinion of the Crowd for rating a movie. This is especially the case when dealing with Twitter data retrieval based on keyword search. The search results invariably include advertisements, retweets, and spam. These have to be filtered out or categorized separately in order to not interfere with the sentiment classification. Since Twitter has a rate limit on data available for use, the workaround for this issue in the application is to store the tweets every time it is retrieved, and calculate a new rating when the movie name is searched again. These results in an increase of data used to review movies, every time the application is used.

5. CONCLUSION

Twitter data effectively manages to capture the opinions and emotions of the crowd and Twitter APIs make it fairly easy to gather this information and analyze it. This desktop application indeed manages to use this massive amount of data to provide a meaningful and useful result due to the speed limit introduced by Twitter, this is often presently an educational implementation of the thought to use twitter knowledge for rating a movie, In future if the limit on the information of twitter is removed then this application has excellent result and may be used for thus several merchandise review and its quality. If all the tweets containing the search string for a movie name will be captured and analyzed, a lot of precise conclusion will be drawn.

The training data used by the classification algorithm are very limited for tweets per category. The results have been encouraging with the small set; I expect the result to be even more impressive with a larger selection of natural language processing in the training data. For some movies we predicted the exact rating but in some cases we are not getting exact result as we have some less data for that movie.

As the application is used frequently, the dataset grows with it. This results in the rating of a movie always being up to date with public opinion. The best use case is to search for recent movies, simply because that is when the crowd seems to be tweeting most about the movies then more accurate rating will be predicted.

6. FUTURE WORK

In future enhancements, more categories can be introduced to classify tweets – extremely positive, mildly positive, extremely negative, mildly negative, neutral, and irrelevant. This can be used to improve the rating formula to make it more accurate. A weight can be associated with each category and then calculate the average.

More number of classes implies increase in number of attributes for the classifier to compare the input with. This should show some difference in the outcome of the classifier and hence difference in the rating shown by the application.

We only consider the data from twitter but in future data from other social media like facebook, YouTube and other blogger comments is also taken into considerations for more accurate rating and with real opinions from all the social media into single place.

7. REFERENCES

[1] Aljaˇz Blatnik, Kaja Jarm, Marko Meˇza, Movie sentiment analysis based on public tweets, Faculty of Electrical Engineering, University of Ljubljana, Trˇzaˇska 25, 1000 Ljubljana, Slovenia, 81(4): 160–166, 2014 ORIGINAL SCIENTIFIC PAPER.

[2] Snehal. A. Mulay, Shrijeet J Joshi, Mohit R Shaha, Hrishikesh V Vibhute, Mahesh P Panaskar, Sentiment Analysis and Opinion Mining With Social Networking for Predicting Box Office Collection of Movie, International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-5, Issue-1)

[3] Ladislav Peska, Peter Vojtas, Hybrid Biased k-NN to Predict Movie Tweets Popularity, Faculty of Mathematics and Physics Charles University in Prague Malostranske namesti 25, Prague, Czech Republic.

[4] Vasu Jain, Prediction of Movie Success using Sentiment Analysis of Tweets Department of Computer Science, University of Southern California,

The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue.

[5] Wernard Schmit, Sander Wubben, Predicting Ratings for New Movie Releases from Twitter Content, Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2015), pages 122–126,Lisboa, Portugal, 17 September, 2015. c 2015 Association for Computational Linguistics.

[6] IMDb.com, Inc. Opening this week. http://www.imdb.com/ (accessed December 2016).

[7] Alexa. Statistics for IMDB, available at

http://www.alexa.com/siteinfo/imdb.com

[8] Alexa. Statistics for Rotten Tomatoes, available at http://www.alexa.com/siteinfo/rottentomatoes.com

[9] Twitter, Inc. Twitter Help Center - Using hashtags on Twitter. https://support.twitter.com/entries/49309 (accessed February 2017)

[10] Twitter, Inc. Discover Twitter - What is Twitter and how to use it. https://discover.twitter.com (accessed February 2017).

[11] Twitter, Inc. About. https://about.twitter.com/company (accessed December, 2016).

[12] Twitter, Inc. Twitter developers - Documentation. https://dev.twitter.com/ (December, 2016)