ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

Mining Twitter UGC to Analyse Box-Office Bombs

Vilas Magare

M. Tech, Department of Computer Science, SGGS Institute of Engineering and Technology, Vishnupuri, Nanded, India.

ABSTRACT: Analysis of user-generated content(UGC) is receiving attention because of its wide application. We address a small subset of UGCmining. Weaddress the UGC related to the biggest Box-Office failures in the history of cinema. We selected 97 movies for the study. We gather related UGC from Twitter, which is a very popular micro-blogging website. Huge data is harvested through a data scraping crawlers. We discuss the sentiment of users on Box-Office failures. We discuss the location and time-zone of different users. we also obtain the association rules from the dataset.

KEYWORDS: Social Media,User generated content, Data Mining, Content Mining.

I. INTRODUCTION

Over the years we saw a substantial increase in the budget and Box-Office collection of movies. But in some cases, the studios had substantial loss over the budget, due to the failure of the movies.

Movies that collect less money from the Box-Office compared to its production cost are considered as Box-Office Bombs. The movie makers have to distribute the gross revenue with the theatre owners, so although the movie makes same as the production cost, still it is in loss. Also, movie makers have to invest in marketing. This also increases the movie cost. A movie has to make more than its production budget if it has to be profitable. The biggest Box-Office bombs are those which are released in summer and has a large competition as many other movies are released at the same time.

There are many reasons for a movie to be unsuccessful. The major cause is heavy competition from the movies which are released in same time, lack of movie promotion, unrealistic production cost, negative word of mouth in social media as well as in the society, negative reviews by the critics, or other external factors such as bad timing or economic problem in the society.

Many A-list directors or actors have at least one Box-Office Bomb. Many suffer major change in their careers after big loss over a movie. Directors such as James Cameron, Michael Bay, Peter Jackson, George Lucas, Steven Spielberg and many more were part of Box-Office Bomb. Actors such as Johnny Depp, Brad Pitt, Julia Roberts, Eddie Murphy and many more were part of Box-Office Bomb. Not only average scripted movie but also classics were part of Box-Office Bomb. Critically acclaimed movies such as" Donnie Darko" (2001)," Fight Club" (1999)," The Big Lebowski" (1998)," The ShawshankRedemption" (1994)," Blade Runner" (1982)," Citizen Kane" (1941)," The Wizard of Oz" (1939) are all been a part of Box-Office Bombs when released. Some movies are unjustly labelled flops, such as Cleopatra (1963) and Waterworld (1995). Some movies are flops in their initial release but gross over international release and further profits by selling it to TV broadcast and home video/DVD release.

In this study, we try to understand the how people reacted to such movies on social media. For this study, we have chosen movies from 1964 to 2016. We have chosen twitter to collect the Electronic Word of Mouth (EWoM), as Twitter is more famous among movie fans. We analyse the tweets of various Twitter users around the world to get the common things among this type of movies. In section II we discuss the related work done on the topic. In section III we discuss the Data collection part. In that we will discuss twitter API, OMDB API and collection of data about Box-Office collection of a movie. In section IV we discuss the experimental methodology we employ in our research. In

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

section V we discuss the Data analysis part. In section VI and VII we discuss the results and the conclusion respectively.

II. RELATED WORK

In this section, we take a brief review of the study in UGCbased analysis on movies. In [1], there is study of movie reviews from a popular on-line database. They randomly sample more than 1000 reviews with titles to train a model. Their model is capable of suggesting words during the process of appraising a film. In [2], a sentiment analysis on movie review has been done. The sentiment profile of a movie is also generated. In [3], researchers applied sentiment analysis and machine learning methods to study the relationship between the on-line reviews of a movie and the movies box office revenue performance. There is very little to no study on movie failures. We try to study this area to find out about biggest movie failures and their reaction on Twitter.

III. PROPOSED METHOD

In this section, we discuss proposed method. Firstly, we discuss getting tweet data, then we discuss getting the information about the movie, in last we discuss getting data about movie box office. In our system, we propose:

a) Data collection module
b) Data cleaning module
c) Content mining module

A. Data collection module: we use several resources like twitter API, OMDB API and Box office mojo for getting our data. Detail description about all is given below.

1. Twitter API: We use Twitter API to get the tweet data. Twitter allows us to interact with its data i.e. tweets and several attributes about tweets, using Twitter API. We need to use a server side scripting language like php, python, ruby etc. to make requests to Twitter, and results would be in JSON format. We use Twitters rest API to get the tweets. Twitter's rest API allows us to get tweets which are stored on Twitter's server. Twitter's rest API is different from its streaming API in the sense that the streaming API allows users to get the live Tweets. In case of live events the streaming API is best option but in our case, it is sufficient to use the rest API, as all our need is to get Tweets about movies which already has been released and labelled as Box-Office failures. Both APIs have their rate limit, upon which we have to rest for few minutes before we start to get the tweet data again. By using rest API, we get the most recent tweet first and the rest follows them. We use python's Tweepy module to authenticate and get the tweets. Python's Tweepy module allows us to authenticate to Twitter by using its OAuth handler. The Tweepy module also allows us to stop when the Twitters limit is reached.

2. OMDB API: The OMDb API is a RESTful web service to obtain movie information, all content and images on the site are contributed andmaintainedbytheusers.Thewebsite http://www.omdbapi.com provides the service. The API gives back results in JSON or in XML format. It requires a key to operate. The key is given by the API providers.

3. Getting Box-Office: info We use the website http://www.boxofficemojo.com to get the box office info about the movies. Box Office Mojo is a website that tracks box office revenue of movies. The website is widely used within the film industry as a source of data. The international section covers the weekly box office of 50 countries and includes historical box office information from three more countries, as well as provides information for box office results for individual films from up to 107. The site also creates an overall weekend chart, combining all box office returns from around the world, excluding the United States and Canada. The overall weekend chart currently tracks the Top 40 films as well as approximately fifty additional films with no ranking.

After getting the data, we store in our database which is a simple comma separated value(csv) file. This data is then processed with the help of different software.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

B. Data cleaning module: We use python for our data cleaning. We use python's NLTK and Textblob modules for cleaning the tweet text part. The details of data cleaning is given in section IV.

C. Content mining module: In this module, we use software such as Microsoft PowerBI, R to mine the content. Fig. 1. Shows our modules.

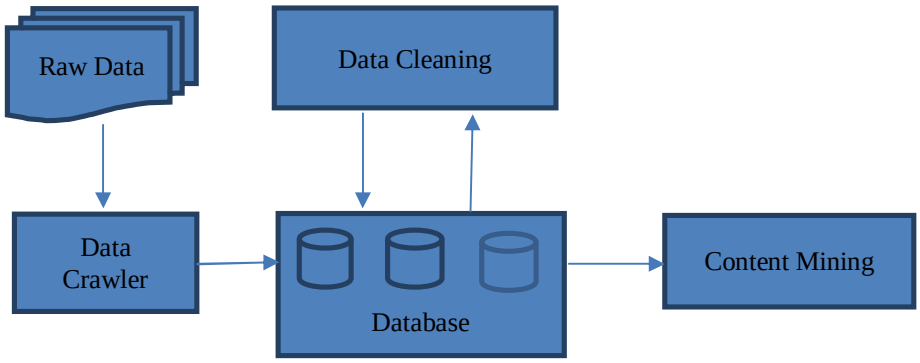

Fig. 1. Architecture of data mining module

Fig. 1. Shows our architecture. In this architecture, we collect the data from twitter using twitter API, with the use for data crawlers and store this data in our database. The data cleaning process will then take place upon the downloaded data. After that apply apriori algorithm on the cleaned data to find out association rules. We also use other software to find out useful information.

We use apriori algorithm t find out the association rules from our dataset. The pseudo code of apriori algorithm is given below.

- *Ck*: Candidate itemset of size k
- *Lk* : frequent itemset of size k
-
- *L1* = {frequent items};
- **for** (*k* = 1; *Lk* !=∅; *k*++) **do begin**
- *Ck+1* = candidates generated from *Lk*;
- **for each** transaction *t* in database do
  - increment the count of all candidates in *Ck+1* that are contained in *t*
- *Lk+1* = candidates in *Ck+1* with min_support
- **end**
- **return** ∪*k Lk*;

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

IV. EXPERIMENTAL METHODOLOGY

In this section, we take a look at the methodology we used in our system. The steps involved in our system are as follows:

1) Data collection
2) Data cleaning

A. Data collection

1) Movie list: We search the web for the biggest Box-office disasters. We get the results and we compare this result to the other source. We then go to boxofficemojo http://www.boxofficemojo.com to get the box office information of those movies. The movies we selected are The 13th Warrior, 47 Ronin, The Adventures of Baron Munchausen, The Adventures of Pluto Nash, The Adventures of Rocky & Bullwinkle, The Alamo, Alexander, Alice Through the Looking Glass, Allied, Aloha, Around the World in 80 Days, The Astronaut's Wife, Ballistic Ecks vs Sever, Battlefield Earth, Battleship, Ben-Hur, The BFG, Blackhat, Catwoman, The Chronicles of Riddick, Conan the Barbarian, Cowboys & Aliens, Cutthroat Island, Deepwater Horizon, Doctor Dolittle, Dudley Do-Right, Ender's Game, Evan Almighty, The Fall of the Roman Empire, Fantastic Four, Fathers' Day, Final Fantasy: The Spirits Within, The Finest Hours, Ghostbusters, Gigli, Gods and Generals, Gods of Egypt, The Good Dinosaur, The Great Raid, Green Lantern, Hard Rain, Hart's War, Heaven's Gate, How Do You Know, Hudson Hawk, Hugo, The Huntsman: Winter's War, Instinct, The Invasion, Ishtar, Jack Frost, Jack the Giant Slayer, John Carter, Jupiter Ascending, K-19: The Widowmaker, Krull, Kubo and the Two Strings, Land of the Lost, The Last Castle, Lolita, The Lone Ranger, The Lovely Bones, Lucky You, The Man from U.N.C.L.E., MarsNeedsMoms,Monkeybone,Nine,TheNutcrackerin3D, The Making of 'One from the Heart', Osmosis Jones, Pete's Dragon, Peter Pan, Pixels, Poseidon, The Postman, R.I.P.D., Red Planet, Revolution, Rise of the Guardians, Rollerball, Rush Hour 3, Sahara, Soldier, A Sound of Thunder, Speed Racer, Sphere, Star Trek Beyond, Stealth, Supernova, Teenage Mutant Ninja Turtles: Out of the Shadows, Titan A.E., Tomorrowland, Town & Country, Treasure Planet, War Dogs, Windtalkers, The Wolfman, xXx: State of the Union, Zoom.

2) Getting data from Twitter: We discussed what we have used to collect data from Twitter in section III. However, we haven't break down the entire process. In this section, we discuss the data collection part in more detail. We use Twitter's API to authenticate and get Tweet data from twitter. We use Twitter's rest API to get the tweets. The rest API allows us to get the Tweet information from the past. The most recent ones will be accessed first and the rest will follow it. Twitter provides every data field information that a tweet has but we don't need every data field. The data fields we need includes tweet ID, tweet time, tweet text, retweet count, favorite count User ID, User's date registered, User's followers, user's verification status, location, time zone.

3) Getting movie info: We collect the movie data from website called as omdbapi http://www.omdbapi.com. The data includes Title, Year, Rated, Released date, Runtime, Genre, Director, Writer, Actors, Plot, Language, Country, Awards, Poster, Ratings (Internet Movie Database, Rotten Tomatoes, Metacritic, Metascore), imdbRating, imdbVotes, imdbID, Type, DVD Release date, BoxOffice, Production House, Website. We collect information about every movie in our listand save it to our database. To get the data from omdbapi we use python. We use python's requests module to send HTTP get request to omdbapi. The omdbapi sends back reply in JSON format. Among the reply we save only the relevant information. The information we save includes Movie name, release year, runtime, directors, actors, writers, IMDB ID, IMDB rating, Metascore, genre, Box-Office collection.

4) Getting movie box-office info: We get the box office info from the website box-office mojo www.boxofficemojo.com/). We manually collect the box office info. We get domestic box-office total, as well as international box-office total. We calculate the inflation if necessary. We also get the DVD release gross. We get Domestic DVD Sales and Domestic Blu-ray Sales and sum them up. Our final is the sum of all.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

B. Data cleaning:

Our collected data contains lot of irrelevant information. Such information may not be useful in our system. Following we show how we clean data from the collected raw data.

1) Tweet data cleaning: When we get data from twitter it contains everything that twitter provides. We are interested in Movie related tweets but when we fire our query, then all tweets containing that query will be downloaded. We remove spam tweets by removing all tweet info which contain link in the tweet text part. We also remove retweets and reply tweets. To remove retweets, we just simply check whether the tweet starts with 'RT @'. For reply tweets, we check whether the tweet starts with'@' symbol. Movie name such as 'Pixels' is a general English word and thus all tweet info containing 'Pixels' in the Tweet text part are downloaded. To remove movie irrelevant Tweets, we train a classifier which separates movie tweets and non-movie tweets. We use discriminative classifier model. The tweets are filtered through this classifier before further processing. After filtering we collect the particular fields of a tweet package. We also add a separate field called as polarity to our database which shows the emotion behind the tweet. To determine the polarity, we have used python textblob, which is a very famous module for natural language processing. We also used NLTK to remove stop words from our tweet text. We then save this optimized data to a csv file.

2) Movie info data cleaning: We use OMDB API t get data information about movies. When we get information about movies, we remove irrelevant fields such as poster, production etc. We then integrate info obtained from box-office mojo to the relevant fields of the info obtained from OMDB API.

V. DATA ANALYSIS

For analytics, we use Microsoft PowerBI. By using Microsoft PowerBI we get useful insights about our data. Microsoft provides quick insights which is a very easy way to find out quick insights from the given dataset. We can also use custom options to find out more. From tweet data, we associate each field with other to find out any correlation or dependencies between them. We also use R's association rule mining algorithm to find out association rules in our data. Weuse R's arules package to find out association rules out from our data.

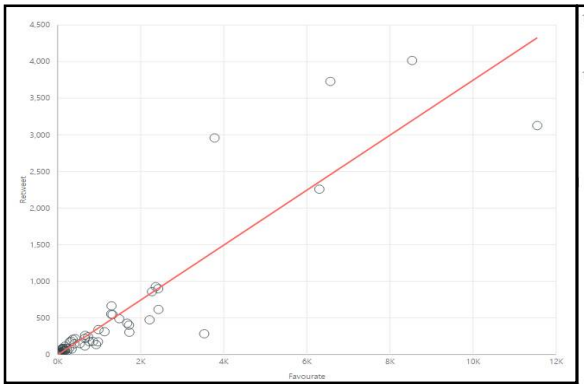

Fig. 2.Retweet vs Favourite

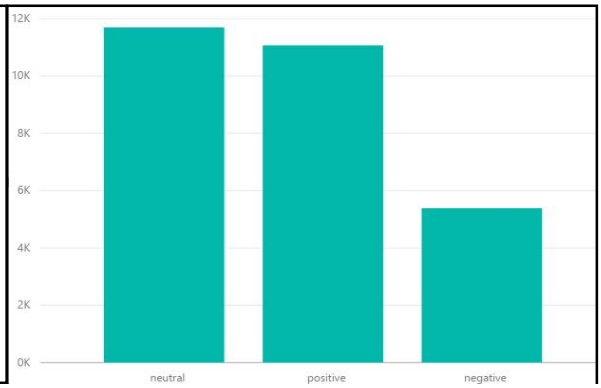

Fig. 3. Retweet by Polarity

Our dataset consists two separate columns for retweet and favorite. The favorite column is the total number of users favorite a particular tweet and the retweet column consist of total number of users retweeted that particular tweet. Fig. 2. shows the co-relation between retweet and favorite. If the tweet is more favorite among users then it is likely that it is retweeted. Our dataset also consists of a column for polarity or sentiment of a tweet. Fig. 3. shows the total number of re-tweets by polarity. The neutral tweets being re-tweeted the most while thenegativetweetsarere-twittedtheleast. Fig. 4. ShowsEastern time (Us and Canada) and Pacific time (US and Canada) have noticeably more followers for polarity 'positive'. Those users having these time zone comprises of more than half of the fan following. The users from time zone London has fairly large fan following for the polarity 'positive' as compared to any single time zone.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

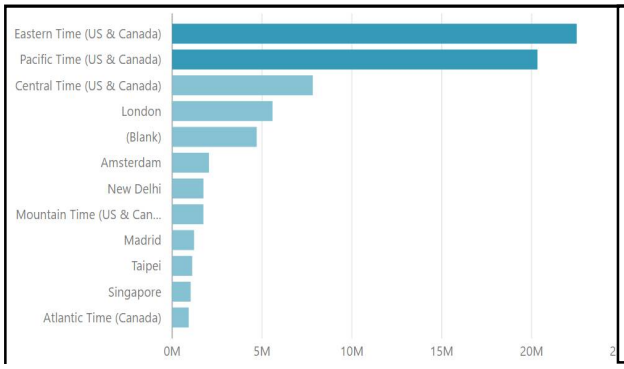

Fig. 4. Time Zone vs followers for polarity positive

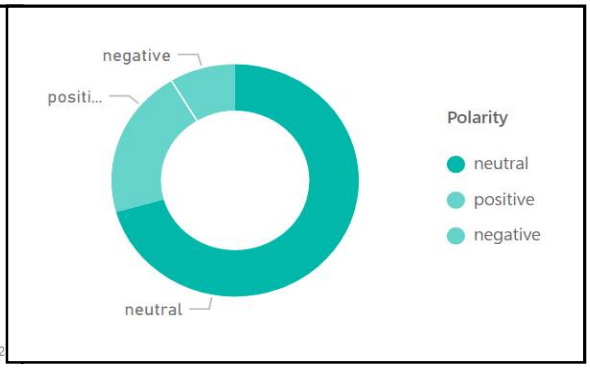

Fig 5. followers by Polarity

Fig. 5. shows that the total count of followers for the users who tweeted with neutral polarity has large number of followers. The follower count for neutral tweets is approximately 70%. while for positive and negative polarity the follower count is approximately 20% and 10% respectively. Table. I. shows the number of users in our data by their time zone. Most of the users do not have particular time zone. Most users are from US and Canada time zone. Fig. 6. shows that 'en' accounts for the majority of thefavorite for polarity 'Positive' Fig. 7.ShowsthatmovieshavingPG-13certificatehavemore gross than movies with other certificate. Fig. 8. shows that among all genre, the action, Adventure, sci-fi has more gross overall. Following frequent user handles has more following. The user handle names are @AngryJoeShow, @DartronRS, @Spazkidin3d, @redsteeze, @AlphaOmegaSin, @dvdinfatuation, @monkeys_robots. These users have many things in common. Firstly, all these users mostly tweet about movies. All of them have followers more than 10000. Lastly all are consistent in their activity on twitter.

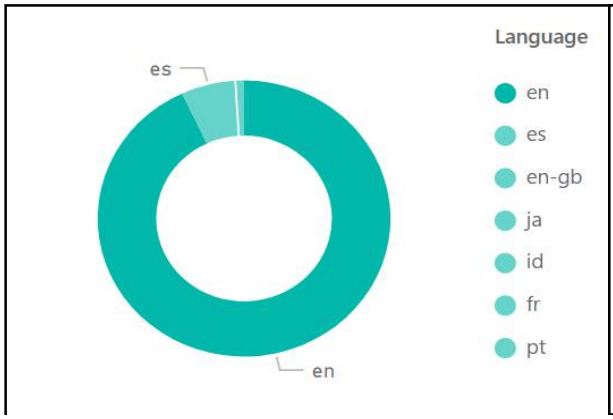

Fig. 6. followers by Language

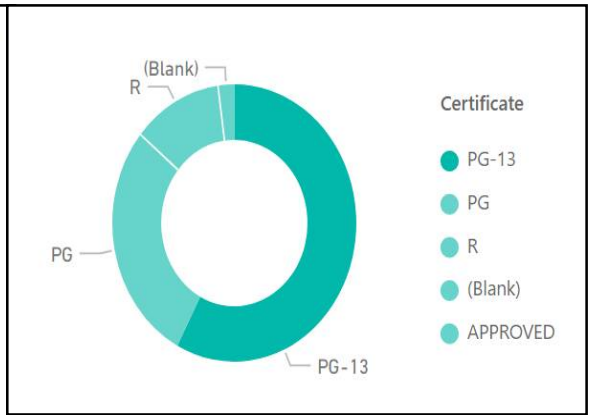

Fig. 7. Gross by Certificate

We also applied apriori algorithm on the dataset as well on the tweet text part of the dataset. The results are shown in Table. II. We see that users with language as English tweeted mostly with positive polarity or neutral polarity. Also, people with language as English have their account as not-verified.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

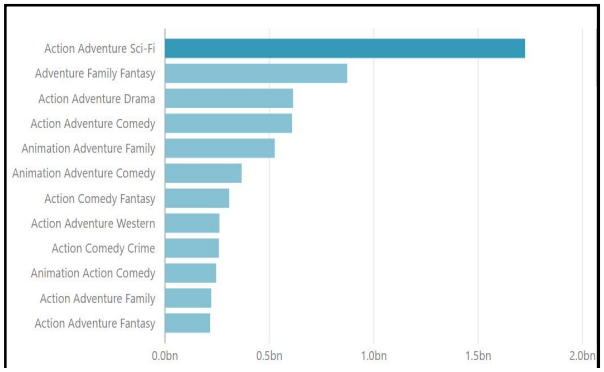

Fig. 8. Gross by Genre

Time zone	number of users
None	12010
Pacific Time (US and Canada)	6046
Eastern Time (US and Canada)	3667
Central Time (US and Canada)	2767
London	1498
Mountain Time (US and Canada)	626

Table. I. Count of users by time zone

Items	Support	Confidence	Lift
Positive => en	0.0104	0.8688	33.4588
Positive => not-verified	0.0117	0.9805	29.3356
neutral => en	0.0111	0.6590	25.3771
neutral => not-verified	0.0166	0.9866	29.5171
en => not-verified	0.0254	0.9796	29.3098
{en, Positive} => not-verified	0.01021	0.9781	29.2642
{ en, not-verified } => Positive	0.01021	0.4013	33.4068
{ en, neutral } => not-verified	0.0109	0.9814	29.3638
quiz => know	0.0052	0.9269	12.2915
xt20launchph => 7daysto	0.0064	1	154.404

Table. II. Association Rules

VI. RESULTS

In this section, we show the useful results among all the results we get from our analysis. The main results we found in our analysis are as follows: The overall polarity of users was neutral though the movies were biggest Box-Office failures. This indicates that the userswere quiet satisfied by such movies. The second largest polarity was positive. This indicates that many people liked such movies. Moreover, the positive polarity is approximately twice that of negative polarity in the entire database. Theretweet count is more for the tweets with neutral polarity. It is slightly less for positive polarity and very less for negative polarity. The users having polarity positive have large number of followers. The users having English as their language on twitter has more followers. This indicate that most twitter users have their language set as English.

VIII. CONCLUSION

In this paper, we mined twitter to find out useful information related to Box-Office bombs. We mined twitter using twitter API. We also gathered the data about the movie itself. We used tools like Microsoft PowerBI, R, and Python to find out important knowledge from our data. We particularly focused on a specific movies that are the biggest failures in the history of cinema. Research in UGC is still at the beginning phase and therefore progress in this area must continue. There are many possibilities of research and improvement in UGC mining.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

REFERENCES

1. Ghosh S., ModakS., and Mondal A. C., "A Model of Opinion Mining for Classifying Movies", International Journal of Advanced Research in Computer Science & Technology, Vol. 3, Issue 1, pp.41-46, 2015.
2. Singh, V.K., Piryani, R., Uddin, A. and Waila, P., "Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification", Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), pp.712 – 717, 22-23 March 2013.
3. Yao R., Chen J., Predicting movie sales revenue using online reviews, Granular Computing (GrC), IEEE International Conference, pp.396 – 401, 13-15 Dec. 2013.