International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2020): 7.803

Revenue Predictor for Hollywood Movies using Sentiment Analysis and Regression

Moraish Kapoor[1], Karshni Mitra[2], Abhilash Arun[3]

Department of Information Technology, SRM Institute of Science & Technology, Chennai - 603203, India

Abstract: ***In today's world, data is being generated faster than ever. A vast amount of this data is freely available on social media. The main goal of data mining and analysis is to utilise this excess, unstructured data to gain insights that an industry can benefit from. The global box office value of the Hollywood movie industry was about $42.2 billion in 2019. Through this project, we aim to predict the revenue of movies on their opening weekend using sentiment analysis on Twitter data to assess the hype around the movie generated by trailers, promos and advertisements. Using this data, along with other features a score is calculated for the movie and used to train a regression model. This model then predicts the opening weekend revenue for upcoming movies. This data can be used by the industry to devise their marketing and release strategy, thereby maximising revenue.***

Keywords: Natural Language Processing (NLP), Machine Learning (ML), User interface (UI), Application Programming Interface (API)

1. Introduction

Revenue generated by a movie depends upon factors other than the quality of the movie. In today's digital world, selling the movie to the people before its actual release has become more viable than ever. The marketing and strategising around the release of the movie is as important as filmmaking itself. The intent of our paper is to enable production houses to better optimise their movie promotion and outreach plan, by estimating the hype generated amongst its target audience.

Social media usage in 2020 was about 3.6 billion people, i. e. a little less than half of the entire population. This figure has increased by about 10% since 2018. Given this steady growth in social media usage, analysis of this data is becoming increasingly important. People use social media platforms such as Twitter, Facebook etc to express their anticipation, excitement or even disappointment about a movie's trailers before the actual release of the film.

In this project, data scraping was implemented by filtering the tweets based on the movie's name and hashtags associated with it. These tweets were then individually processed to calculate a polarity score ranging between - 1 and 1 which signified the user's sentiment towards the movie; - 1 being a negative tweet and 1 being a positive one. These polarity scores were then averaged to get an overall estimation of the public sentiment towards the movie. Using this feature, along with features like lead actor, movie genre, number of tweets and screenings, a database was created that the regression model utilized for training.

The model identified the significance of the various features and enabled us to fine tune it by assigning more weightage to those features therefore generating optimum results. Additionally, the database is updated with new weekly releases and the model is retrained at fixed intervals to improve future predictions.

2. Literature Review

2.1 "Business Intelligence from Social Media: A Study from the VAST Box Office Challenge"

The paper uses a model which is based on the user's interaction with the system and not on a model that is based purely on math. This interactive behaviour can be extended to several domains where it may be useful such as, advertisement, sales, finance etc. However, this analysis is not specifically done on the movie industry.

Algorithms: Multiple Regression Model (Linear), Prediction Temporal and Modeling Unsupervised Layer.

Drawbacks: Social media data is extremely noisy due to the large amounts of users and different social media platforms. Due to this, any product to perform such predictions would eventually be steered off course. To keep such a system working, regular updating is required to meet the current social media standards and requirements.

2.2 "Sentiment Analysis for Social Media"

The author extracts features from the text generated by the user and this is then fed to NLP algorithms. Doing this allows the author to extract features from large volumes of text data and therefore, obtain the public sentiment of the users. Dictionaries are made which map the words and therefore we get the polarity score of the text.

Algorithm: Naive Bayes Classifier

Drawbacks: The sentiment score is solely based on certain words and not on the context in which they are taken.

2.3 "Predicting Box Office Revenue for Movies"

The project first extracts feature from the Movie database present on IMDB. The paper then uses these features to build a prediction model which gives better results compared to other existing models. The algorithm used is unsupervised clustering. This algorithm does not have a fixed output and it

analyzes the data and tries to classify it into clusters which show the most similarity; thereby giving the collected data some structure. The goal here is to get some meaningful understanding of the data.

Algorithm: Unsupervised Clustering

Drawback: The algorithm used in this paper clusters movies based on similarities and does not take into account the public sentiment to optimise revenue.

3. Proposed System

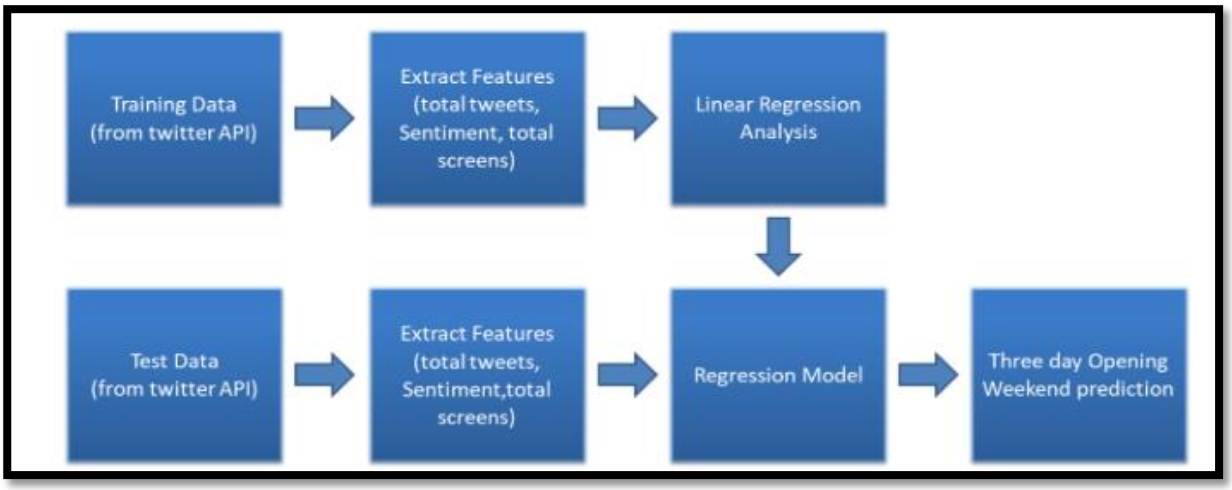

Figure 3.1: Architecture diagram

3.1 Module 1 - Data Collection

Twitter is used as a source to gather data about movies as it is a platform where a significant proportion of the global population expresses their opinions on a variety of subjects. A web scraper is used to collect the required Twitter data based on a specific date range and select keywords such as movie name and other relevant hash tags.

3.2 Module 2 - Sentiment Analysis

Analysis of a tweet is done to identify if the sentiment behind it is positive or negative. This is accomplished using a python library – 'TextBlob' which processes each tweet and assigns it a polarity score. TextBlob uses a dictionary which contains a list of words and their respective sentiment values, and based on this we obtain a final score for each individual tweet.

3.3 Module 3 – Regression

This processed data is then used to train a ML based regression model. The model takes in the polarity score, number of tweets, screenings and the movie genre and uses this information to predict the opening weekend revenues of the upcoming movies.

3.4 Module 4 - User Interface

A web application is created using HTML, CSS, SQL, PHP and JavaScript to allow the user to select any upcoming movie and obtain its predicted revenue. A historical report of past predictions is also accessible to the user for reference.

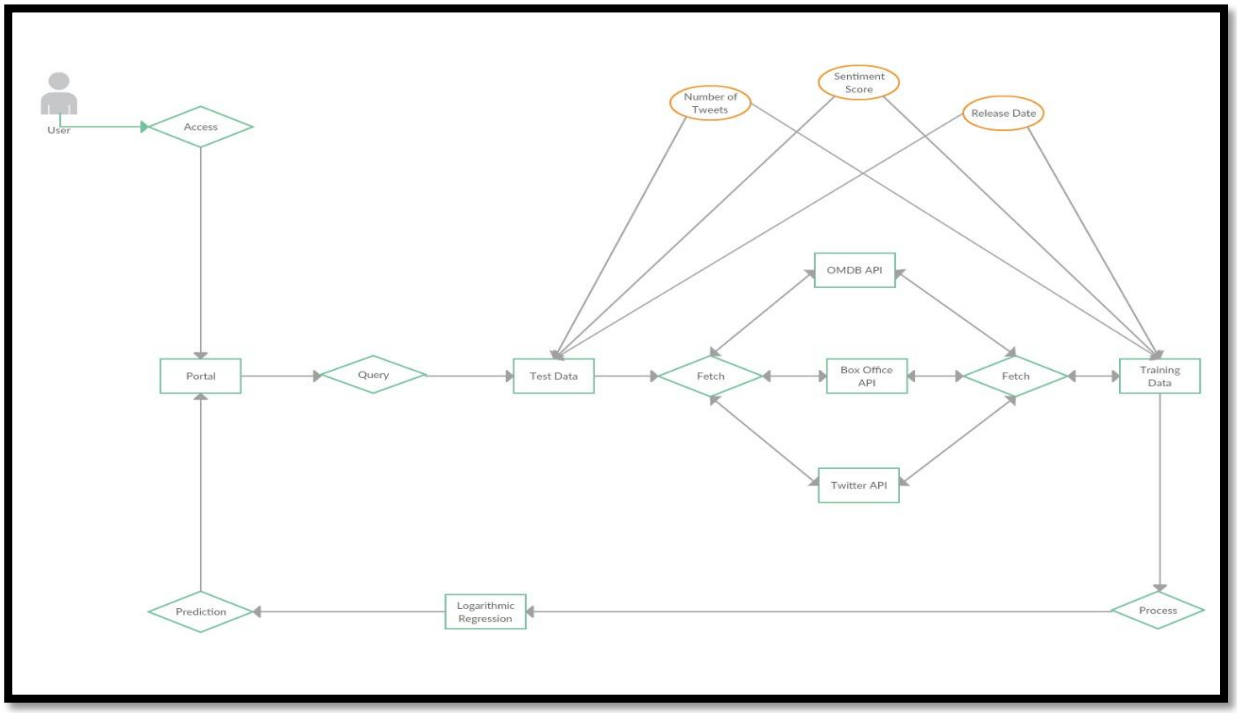

Figure 3.2: Entity relationship diagram

4. Implementation methodology

4.1 Module 1: Collection and extraction of data

Every second, around 6000 tweets are posted by Twitter users. This huge amount of data is unstructured and unfiltered. However, if analyzed, valuable information can be extracted. 'GetOldTweets3' is the python library used in the project to collect data. It is essentially a web scraper that scans Twitter based on certain parameters. The scraper traverses each page just like a user would and fetches the required data. It utilises Twitter's built - in search functionality by passing in the given parameters that is, the date range and the keywords. Then, it collects all the tweets satisfying these criteria and stores it in a file. The user also has the option to download the tweets for a given movie using the web interface.

Additionally, we also extract data on the number of screenings of the movie. This plays a critical role as the revenue is always limited by the number of seats available and screenings of the movie. This information is extracted using the Box Office Mojo API. The API also provides additional features such as the cast and genre of the movie which then act as input features for the regression model.

Table 4.1: Database Table containing test movie data, their release date and hashtag

1	Name	Rel_Date (dd - mm - yyyy)	Hashtag
2	Antman And The Wasp	04 - 07 - 2018	#AntManAndTheWasp
3	Bad Times At The El Royale	11 - 10 - 2018	#BadTimesAtTheElRoyale
4	BlacKkKlansman	09 - 08 - 2018	#BlacKkKlansman
5	Venom	05 - 10 - 2018	#venom
6	Bohemian Rhapsody	24 - 10 - 2018	#BohemianRhapsody
7	Crazy Rich Asians	15 - 08 - 2018	#CrazyRichAsians
8	Fantastic Beasts	16 - 11 - 2018	#FantasticBeasts
9	Game Night	22 - 02 - 2018	#gamenight
10	Mission Impossible Fallout	12 - 07 - 2018	#MissionImpossibleFallout
11	Ready Player One	28 - 03 - 2018	#readyplayerone
12	The Grinch	22 - 10 - 2018	#TheGrinch
13	The Meg	10 - 08 - 2018	#TheMeg
14	The Nun	07 - 09 - 2018	#TheNun
15	Batman v Superman	25 - 03 - 2016	#BatmanvSuperman
16	Captain America Civil War	06 - 05 - 2016	#CaptainAmericaCivilWar
17	Deadpool	02 - 12 - 2016	#deadpool
18	Deepwater Horizon	30 - 09 - 2016	#Deepwaterhorizon
19	Doctor Strange	04 - 11 - 2016	#DoctorStrange
20	Sully	09 - 09 - 2016	#sully
21	Ghostbusters	15 - 07 - 2016	#ghostbusters
22	Hacksaw Ridge	06 - 11 - 2016	#Hacksawridge
23	Jason Bourne	29 - 07 - 2016	#JasonBourne
24	Jungle Book	15 - 04 - 2016	#junglebook

25	Kungfu Panda	29 - 01 - 2016	#Kungfupanda
26	Moana	23 - 11 - 2016	#moana
27	Dunkirk	21 - 07 - 2017	#Dunkirk
28	Wonder Woman	02 - 06 - 2017	#Wonderwoman
29	Thor: Ragnarok	03 - 11 - 2017	ThorRagnarok
30	The Lego Batman Movie	10 - 02 - 2017	#LEGOBatmanMovie
31	Get Out	24 - 02 - 2017	#GetOut
32	Baby Driver	28 - 06 - 2017	#BabyDriver
33	Coco	22 - 11 - 2017	#PixarCoco
34	Justice League	17 - 11 - 2017	#JusticeLeague
35	American Made	29 - 09 - 2017	#AmericanMade

4.2 Module 2: Sentiment Analysis

The tweets collected in the previous module are then analysed and assigned a polarity score ranging from - 1 to 1. Every tweet collected is analysed individually and assigned a sentiment score depending on the content of the tweet. The sentiment analysis module scans the tweet for keywords that indicate whether the tweet is positive or negative.

Figure 4.2 Sentiment scoring workflow

The words that the module is unable to score are automatically assigned a score of 0. For example slang, a misspelt word, words in other languages etc. This is the reason our systemis restricted to movies that have a significant number of English tweets. The module even detects and assigns appropriate scores to emoticons such as "☺" and "☹".

After each tweet is scored, the total is calculated and is divided by the total number of tweets processed, giving us an average polarity score for the movie.

Additionally, the number of tweets indicate the number of people talking about the movie before its release providing an idea about its reach and the hype around it.

4.3 Module 3: Regression

The data collected is then processed by a regression model to estimate the revenue. During model training, tweets collected are plotted against the actual revenue generated by the movie. Then a line is fit through all these points minimizing the sum of the distances of all these points from the line. This is the best fit line. Extending this line to the x - axis we get the x - intercept. After saving this x - intercept, another graph is plotted based onthe number of screens and the actual revenue generated. The x - intercept is stored again. A similar procedure is following for the genre, cast and number of tweets.
These values are then put in the linear regression equation which is, y = M1X1 + M2X2 +. . +MiXi + C

Here y is the expected opening weekend revenue. M1, M2 … Mi are the coefficients calculated. Based on these values a final regression line is fitted through the multi - dimensional graph.

After substituting the values in the equation, we obtain the value of Y which is our predicted opening weekend revenue for the given movie.

After the release of the movie when the actual revenue is available, we calculate the accuracy of the algorithm using |y' - y |, where y' is the actual revenue, and y is the predicted revenue for the movie. This absolute value is then added to the database and the historical prediction report is updated.

This result is then displayed to the user on the UI of the web application. The model is retrained regularly with the updated data of the newly released movies to optimise revenue prediction in the future.

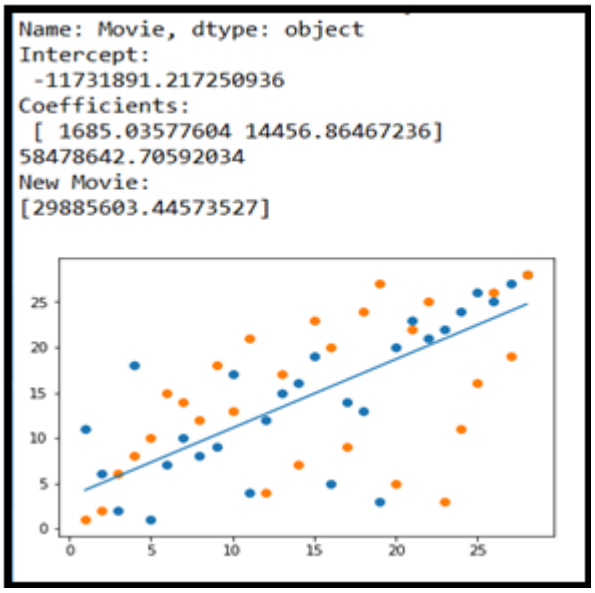

Figure 4.3: Fitting the Regression Line

4.4 Module 4: User Interface and Database

After regression is completed, the result is displayed to the user in a comprehensible format. To create the website HTML, CSS and JavaScript are leveraged for the frontend design and PHP for the backend logic. The database handling is done using SQL.

The user is presented with a drop - down menu which is populated with the upcoming movies. The user then selects the movie they are interested in, and a summary of the movie along with our predicted value is displayed. To make the module transparent, features such as number of tweets collected for the movie, the polarity score and the number of screens that are showing the movie are also displayed.

Additionally, the raw tweets collected for the movie can also be viewed on a separate page of the webapp.

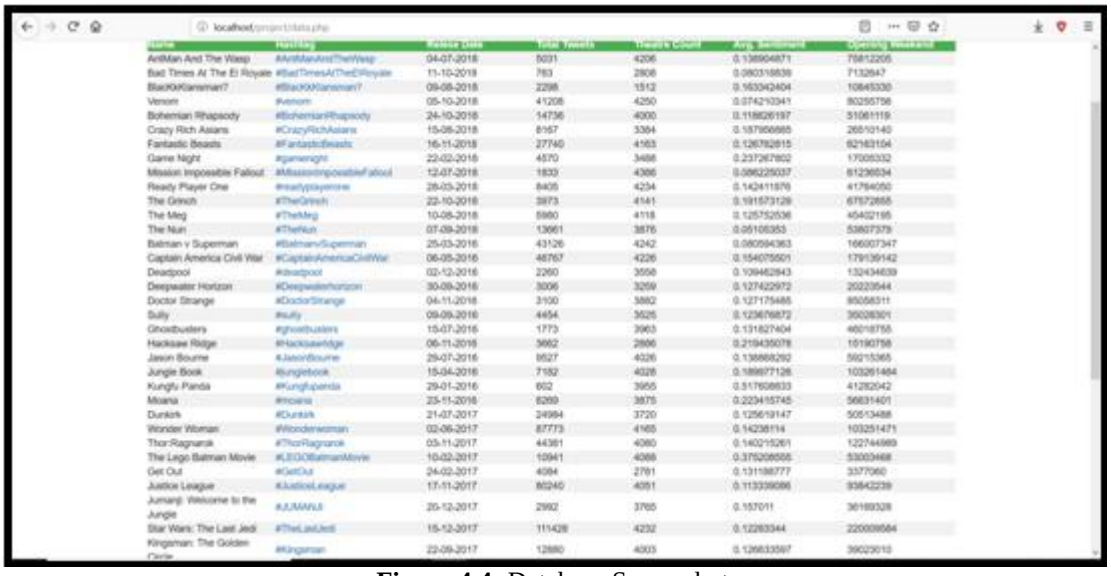

Figure 4.4: Database Screenshot

Figure 4.5: Homepage Screenshot

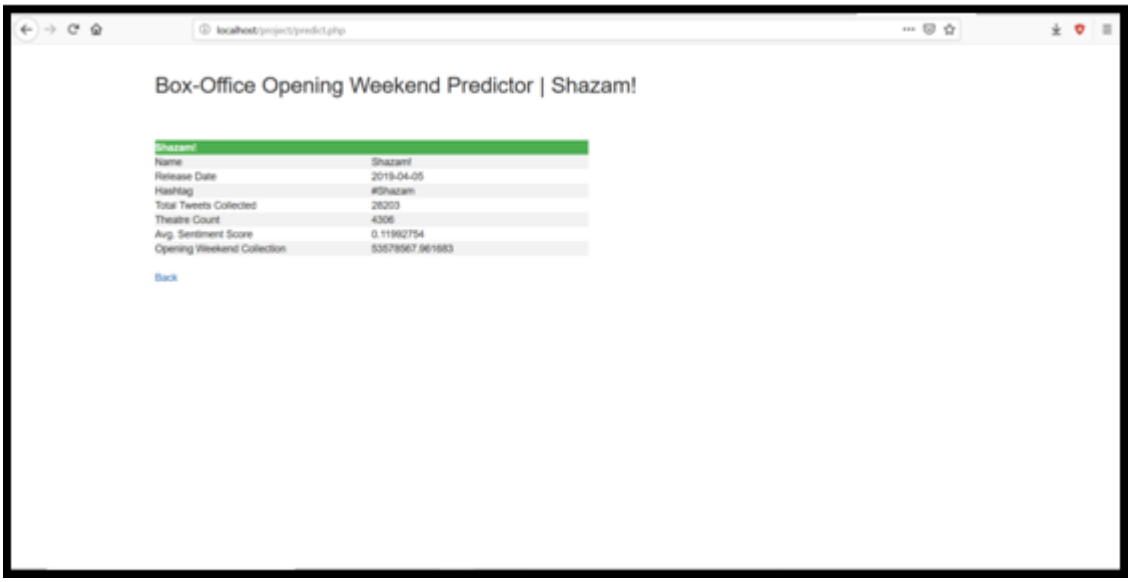

Figure 4.6: Result Screenshot

5. Conclusion

Through this project, we extensively analysed a large amount of unstructured social media data to gain valuable business insights. The training set was built on data collected from older movies from 2016, 2017 and 2018. Approximately800, 000 tweets were collected and analysed.

The 'pass criterion' of the revenue prediction for a movie is defined as – if the predicted value lies within a range of 10% of the actual revenue. The system was able to predict the opening weekend revenues with an accuracy of 87%. This implies 87% of the predicted revenues fall in the range of + - 10% of the actual results.

The bar graph below depicts the predicted opening weekend value against the actual opening weekend value for a randomly selected subset of the data.

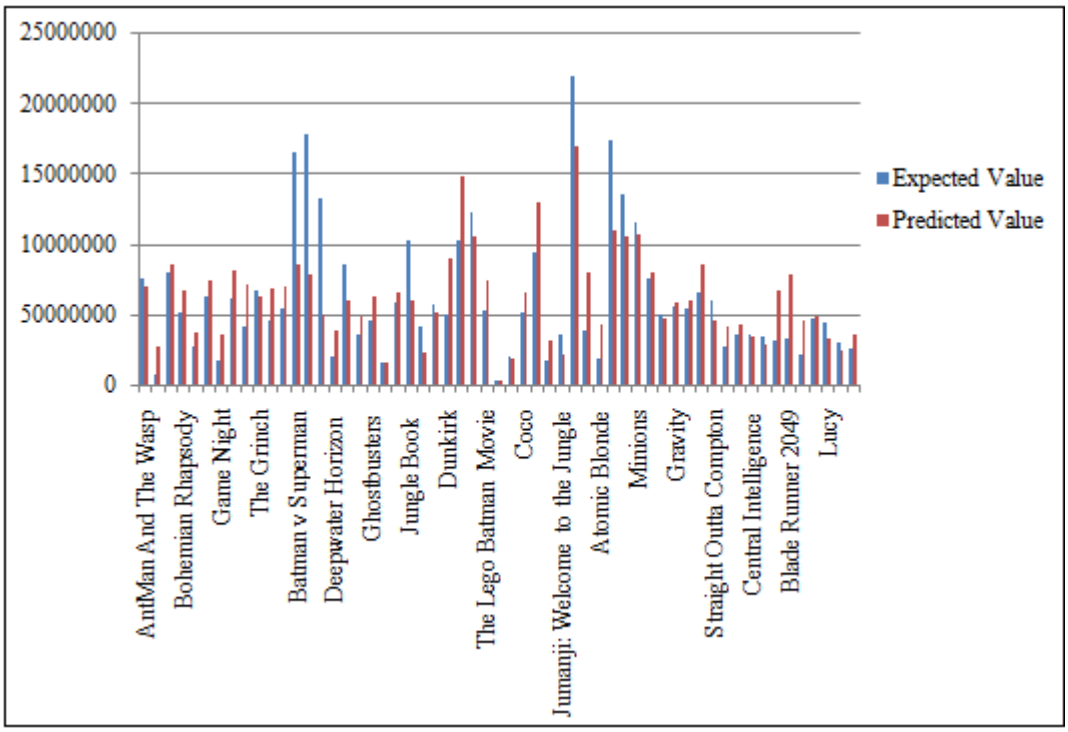

Figure 5.1: Bar graph depicting Expected vs Predicted Revenue

6. Acknowledgement

The authors would like to thank Ms V Vijayalaxmi, Assistant Professor, Department of Information Technology, SRMIST, for her sincere efforts in helping and guiding through the research paper. Her insights and suggestions have helped in greatly shaping the project.

References

[1] Afeng Lu, Feng Wang, and Ross Maciejewski. "Business Intelligence from Social Media: A Study from the VAST Box Office Challenge". IEEE Computer Graphics and Applications, Volume: 34, Issue: 5, Sept. - Oct.2014.
[2] Aida Mustapha, Farhana M. Fadzil, "A Regression Approach for Forecasting Vendor Revenue in Telecommunication Industries": Aida Mustapha et al. / International Journal of Engineering and Technology (IJET).
[3] Dr. Md. Haider Ali, "A Machine Learning Approach to Predict Movie Box Office Success": 20th International Conference on Computer and Information Technology (ICCIT 2017).
[4] Matt Vitelli, "Predicting Box Office Revenue for Movies"
[5] Ting Liu & Xiao Ding &Yiheng Chen, "Predicting movie Box office revenues by exploiting large - scale social media content. ": Multimedia Tools and Applications February 2016, Volume 75, Issue 3.