Movie Evaluator

(ICSI: 531 Data Mining)

Dhruv Patel (001313337)
M.S. in Computer Science
University at Albany, SUNY
Albany, New York 12222
Email: dmpatel@albany.edu

Tejas Shah (001311049)
M.S. in Computer Science
University at Albany, SUNY
Albany, New York 12222
Email: tshah@albany.edu

Kushal Patel (001311990)
M.S. in Computer Science
University at Albany, SUNY
Albany, New York 12222
Email: kppatel@albany.edu

Abstract - **Social media is a source of rich information about people's opinion liking and disliking on things. On a microblogging web platform like Twitter people share, support and post their thoughts, opinion and reviews about situations or an occurrence of an event. Twitter generating huge amount of data on daily basis. Twitter data explores different review on same occurrence. In this study we collect tweets, based on the collected twitter data, analyze IMDB ratings and based on that we will try to predict success of movie the main result we will present are whether a movie would be successful at the box office.**

Keywords:

SVM (Support Vector Machine*)*, Twitter rating from IMDB (Internet Movie Database), Tweet classification algorithm, Movie rating algorithm, Movie popularity prediction.

I. INTRODUCTION

In this work, we are trying to predict which movie will gain more popularity in compare to other movies which are also released on same week. We collected tweets from Twitter. Twitter is providing a social networking service and microblogging where users post, share, react, review, give an opinion on incident or an event by interacting through messages and "tweets" which are restricted to 140 characters. Authenticated and Registered users can post tweets, share media on it. But those who are unregistered can only read them. Twitter can be used via web application, SMS or Mobile Application. Twitter Inc. is based in San Francisco, California, United States, and has more than 25 offices worldwide. We are using Twitter Dataset as a part of this work and we collect tweets using Tweeter Rest API. Tweeter API can be used after signed in with authenticated credentials of Id and password. Tweets contains tweet text and unique tweet id as a basic attribute with tweet. We collected tweets of movies are already released and which will be going to release soon.

We gave ratings based on IMDB ratings to released movie. IMDB is well known for Internet Movie Database which provide ratings to released movie. IMDB database includes additional categories like filmmakers, biographies, and plot summaries. The movie ratings contained in a data list. Authenticated and registered users can also submit reviews of movies and TV shows based on 1 to 10 scales, mean of total rating given by all users shown with the name of Movie or TV season. Also, based on the IMDB provide top list of TV seasons and movies.[5]

For our approach, we collected almost 3000 tweets as a part of this work. Our training data set contains movie which are already released and ratings given by IMDB. Our testing data set contains movies which are going to release in near future. We used Data Mining approaches to predict Box Office results by classifying the movie as three categories: Hit, Flop, and Average. [1]

II. Related Work

In Recent years using social media data and giving prediction based on behavior of public thought and reaction is a popular. Nowadays lots of work have done based on this concept as part of Artificial Intelligence, Data Mining and Machine Learning era. Every social media generating lots of data and from that amount of data we can made analysis of public behavior. Using sentimental analysis of public review is common stuff in any of these 3-field's approach. In previously introduced work they additionally introduce an added sentiment category - the neutral category. In this type, with the use of Python programming and NLTK library and compare so obtained results with the normal machine learning. Predicting Ratings for New Movie Releases from Twitter users actively expressing themselves on-line, a large quantity of information is generated daily. Nowadays twitter based sentimental analysis gives more

accurate review of movie rating rather than previous market analysis. Sentimental analysis of twitter data is also hot topic. [2][3]

However, very less work has been done directly with the use of sentiment analysis and get results to predict the future with the use of sentimental analysis. But did not do sentiment analysis to predict the movie success. [4]

III. Body of Paper

The body of the report contains the information about the proposed approach, dataset, how the calculations performed on the dataset and the details of the result we got after the calculations and implementing a methodology. The report ends with the conclusion that is reached after the calculations performed on the dataset based on described approach below.

A. *Proposed Approach*

Our project is focused on, to predict which movie is going to be more successful based on the current movie gossip on tweeter, based on ratings provided by IMDB. We tried to predict a rating of upcoming movies based on which movie is praised publicly on twitter and by comparing that we come to know which movie gain more popularity.

B. *Methodology*

I. *Data Collection*

Two approaches used for data collection from twitter. With help of search rest API and Steam API to access Twitter public data, API is required that is provided by the Twitter. To access twitter data using this Twitter API it required authenticated user Id and password. Twitter allow access of public with after applying valid authentication key. [2][3]

With the help of this search Twitter rest API we can fetch public tweets but there are certain limitations of API that we can send only **180 Requests per 15 mins window**. per request you can ask for maximum **100** tweets, giving you a grand total limit of **18,000 tweets/15 mins**, if you download 18K tweets before 15 mins, you won't be able to get any more results until your 15-min. window expires and you search again. Also, you need to be aware of the following limitations of the search API. We search movie related tweets based on geolocation when using this approach of collection, a tweet. And run code manually in certain gap of time. [6]

We also used another API known as Steam API also provided by Twitter. With the help of Tweepy we got tweets useful tweets for our task. In this API, we can continuously retrieve tweets form twitter. This API, continuously retrieves data from the global stream of Tweets data. In python code, we need to give keyword to search for and API starts searching on keyword with random geolocations. We used different movie names to get useful data. And output is in Jason format that contains

- Tweet Id
- Username of person who tweeted
- Tweet text
- Time of tweet
- The method the user sends the tweets (e.g. iPhone, Android, etc.)

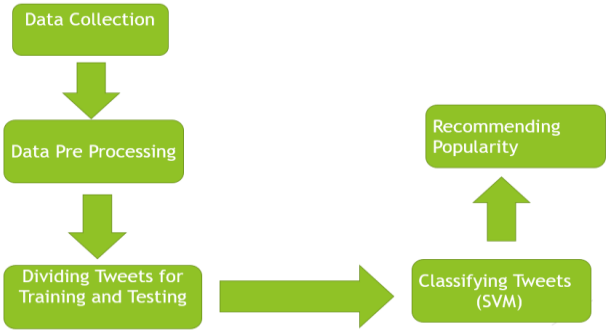
The collected data is stored as a text file for each movie. The data fields are separated by tab [6]

II. *Dataset Preprocessing*

We collected the tweets based on shown Data Collection method. Data collection was done properly in json format but there lots of issue with collected data in terms of duplicate tweets, missing value of any attributes, tweets in different language. So, to avoid that we did data processing. Steps taken on data as Data Processing. [2][3][4]

1) We collect data in single file for all movies so for easy analysis and data processing we created different file for each movie with the help of python code.
2) For our work, we need only data related to tweet text only so rest of the fields like tweet Id, tweet time etc. we removed with the help of python code.
3) For Removing redundancy of data or tweets we did python coded that compare each tweet with all collected tweet in file and write a unique tweet in new file.
4) For tweets, we collected in different language because we used global data steam tweets. we did changes in code that changes languages to English.

After done with data processing we got relevant data.



III. Dividing Tweets for training and testing

To setting up Training and Testing dataset for our work. As a part of training set we use tweets of 15 movies which are already released. We use around 2000 tweets out 3000 tweets as a part of training data. We also collected movie related tweets which are going to release in near future. For that purpose, we use 5 movies as a part of testing dataset. Movies we took as a Training dataset we did further manual data processing. We add IMDB rating with that. With doing this we will get keywords for testing movie data set. Shown image is a list of movies taken under Training dataset. [3][4][9]

Training table:

MovieNumber	Movie Name	Rating
1	It comes at night	2
2	John Wick	4
3	Get Out	4
4	How to Latin Lover	4
5	xXx	5
6	Smurfs	3
7	Kong	2
8	Logan	4
9	Lord of the Rings	4
10	Gifted	3
11	Everything	3
12	Baby Driver	3
13	Atomic Blond	3
14	The pirates of Caribbean	4
15	La La Land	5

In image, there are 3 columns one showing movie names and rating column showing rating based on IMDB. In our work, we took shown movies as a part of testing data. we implemented SVM on that and try to predict rating.

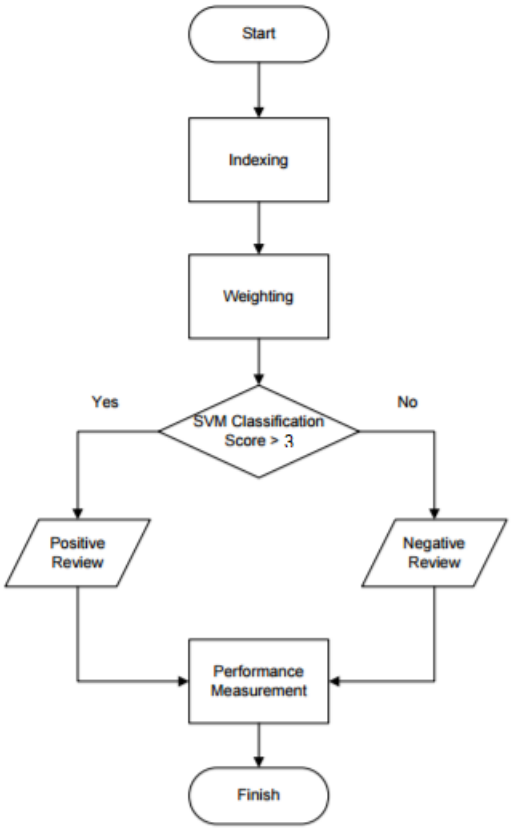
Testing Table:

Number	Movie Name
1	Guardian of Galaxy
2	Baywatch
3	The Boss Baby
4	Bahubali2
5	King Arthur

IV. Classifying Tweets

For our dataset, we can use multiple classification model such as Logical Regression, Linear Regression, Support Vector Machine, Multi Class Regression. Accuracy of prediction is based on which classification model we used. Here we are using Support Vector Machine(SVM) classification model. SVM is well known as best supervised learning classification algorithm.

SVM approach we used:



SVM approach

Show diagram is giving clear idea of implementation of our approach.

Based on Training dataset of 15 movies it will find the feature set and based on that match we give that a particular class label that gives a rating to testing movies.

Step 1: Indexing: In this phase, we are matching our testing dataset to training dataset index by index.
Step 2: Weighting: In this phase, we do weighted min featured word count for particular testing data.
Step 3: SVM classification: In this phase count will be compared to ratings if count is greater than 3 out of 5 than it will be considered as positive review else it will be considered as negative review.
Step 4: Performance Measurement: Final Output of all 5-testing dataset will be in descending order.

Testing table is a showing an output of our work.

Testing table:

Movie Number	Movie Name	Rating after SVM
1	Bahubali 2	4
2	BayWatch	3
3	The Boss Baby	3
4	Guardian of Galaxy	3
5	King Arthur	3

IV. Conclusion and Future work

The results of our work show that we can predict rating of upcoming movie and popularity gain by movie can be predict precisely by sentiment analysis of the movies with the use of simple metrics and prediction of that ratings are with good accuracy. And Predictions are very accurate and near to real box office reviews. We understand that there might be more than one factor which affect the movie box office success, we feel that there is a long list of future work. However, this problem itself is an interesting and promising area.

Some bottlenecks we faced were:

- Collecting unique tweets using rest API.
- It's a time-consuming process to collect huge amount unique twitter data from different geolocation throughout world. More data we got more accurate prediction will be.
- Language Barrier in tweet because of global geo locations

Future Work: Same things can be done with the help of different classification or clustering models which may provide better prediction of popularity gain by movies.

V. References

1. Wikimedia Foundation
https://www.wikipedia.org/

2. Research Paper on Prediction of Movie Success using Sentiment Analysis of Tweets by Vasu Jain
http://jscse.org/papers/vol3.no3/vol3.no3.46.pdf

3. Research Paper on movie rating prediction based on twitter using sentiment analysis by Mr. Abhishek Kesharwani, and Mr. Rakesh Bharti
http://www.jacotech.org/uploads/1488123490__64067899.pdf

4. research Paper on Predicting Movie Ratings of IMDB Users by Jingqiu Zhou,Mingyuan Xiao,Xiaoguang Mo
http://www.cse.scu.edu/~mwang2/projects
Predict_movieRating_16w.pdf

5. IMDb.com, Inc. An Amazon.com company.
http://www.imdb.com/

6. Twitter, Inc.
https://twitter.com/

7. Python Software Foundation
https://www.python.org/

8. Slide refernce by Vasu Jain
https://www.slideshare.net/vasujain/sentiment-analysis-of-tweets

9. Lecture Notes on SVM model by Andrew ng
http://cs229.stanford.edu/notes/cs229-notes3.pdf