MovieBuzz : Deriving A Formula To Make Successful Films

Apoorva Walimbe
Computer Science
New York, USA
acw458@nyu.edu

Aastha Mahajan
Computer Science
New York, USA
am6621@nyu.edu

Rajat Pawar
Computer Science
New York, USA
rp1978@nyu.edu

Abstract—

With recent advancement in technology, people now have numerous media to voice their opinion. There are many factors responsible to make or a break a product. All the industries are facing positive as well as negative effects due to this and entertainment industry is no exception. Our research combine the classical factors behind predicting success of a movie with these new-age technical factors. We aim to design an analytic to define a correlation between various factors and come up with a formula to determine how to make a successful film.

Kewords—

Predictive models, Entertainment industry, Twitter, IMDB, Success prediction, Trend analysis, Highest grossing movie

I. Introduction

The project MovieBuzz will concentrate on analyzing current film trends and deriving a success formula from the data that is available to us. Our main aim is to analyze the current trends in movies. By trends we mean, what is appreciated by the audience. Audience react to various stimuli nowadays and it is difficult to say whether your work would be a hit or a flop. With our project, we will aim to design one such success formula for upcoming projects. Every single day there is a new thing that's trending in the market. We will collect real time data of such events and will tell the entertainment industry how they can benefit out of it. The top 25 lists of every kind alters almost every other day. We will analyze the data from IMDB datasource which will give us an insight as to what is appreciated by users. We will combine it with the real time data analytics of twitter data. At the end of the analysis, we will deduce which artists are most likely to give a hit film paired with which director and if possible, what should be the genre of the film and the plot of the film. We hope to come up with a relation that binds it all together and leads to successful future prediction.

II. Motivation

One of the biggest challenges that most movie studios and filmakers face is how well will their next movie do in the box office? Up till now, the answer to this question has been a lot of vague guesswork. But with the emergence of social media sites, as well as big data technologies, movie makers now have a way to measure sentiment of their audience by accessing multiple big data sources.

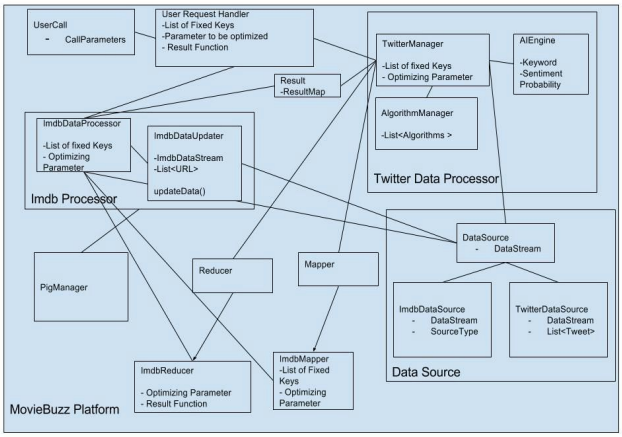
In past few years, social media has become ubiquitous and crucial for social networking and content sharing. Still, the data that is generated from these sources remain largely untapped. The primary motivation behind MovieBuzz was to make an analytics that could provide meaningful predictions on various aspects that are critical to a movie's commercial success.

III. Related Work

Everyone now realizes the importance of social stimuli in predicting audience's reaction to every product. A lot of research is going on relating to how social factors are taking over classical predictive anlysis factors. We have analyzed scholarly papers related to our field of research. During the research, we came across many theories which try to define the correlation between classical and social factors.
We analyzed the papers which describe the current models in detail and show how each factor contributes to these models. One of the papers show how critic's reviews of the film plot affect the first weekend revenues of the film.

The problem with most movie recommendation algorithms is to estimate whether a user would like a movie he hasn't seen, based on the response of 'similar users' for that movie. This can be considered as figuring out the missing values in the user-movie matrix. One more research paper suggested a memory-based collaborative filtering approach for prediction and recommending new movies to users.

IV. DESIGN



Tentative Level 0 class Diagram of Data Processor for moviebuzz

We are making use of multiple datasets to make a final prediction based on our analytic in the project. The IMDB dataset is a historic dataset which contains information about movies released in past. It contains various attributes like actor, actress, genre of the film, rating, etc. This dataset makes use of a (key,value) model. We have a "fixed set" of key value pairs and one "parameter(key, herein)" to be optimized. This parameter is the key on which data will be grouped in the "reduce" phase. Eg. (very simple) For a problem where we want to know the tentative revenue of a film produced with certain parameters (like actor, director, actress and genre) :

Fixed Set = [(Actor, {Alex Baldwin}) , (Director, {Woody Allen}), (Actress, {Kate Winslet}) ,(Genre,{Crime})]

Parameter(to be calculated): Revenue

Static data from Imdb contains data related to various parameters in normalized text files.

We also make use of real-time data streamed from Twitter. The real-time data gives us information about people's sentiment. Twitter users post their reactions on the platform with the help of hashtags. We filter the live data based on these hastags and then perform sentiment analysis on them. The user revies are then categorized based on the results we get from sentiment analyzer.

The metadata gathered from processing of these two datasets is then combined to make a final prediction about the success rate of the movie.

V. RESULTS

We streamed data for 15 movies from Twitter. We started cleaning the data as soon as we had gathered sizeable amount of it. We observed we're getting a lot of redundant data because of the retweet features. The streaming API doesn't provide with a default functionality to ignore retweets. Hence, we decided to manually ignore the retweets. We modified the code accordingly and started gathering data again.

It was important to analyze the sentiment behind each tweet in order to know what the audience have to say about these upcoming movies. There are many ways to perform sentiment analysis on streamed tweets. We decided to use Alchemy API for sentiment analysis. After modification of the code, we had the file ready with the tweet and its sentiment. We further modified the code to analyze the tweet immediately after streaming and adding it to the file and continued gathering data from Twitter. Alchemy API tells you if the tweet is either positive or negative or neautral. By analyzing the ratio of positive to negative to neautral tweets, we got the most anticipated movie according to the audience.

VI. FUTURE WORK

The social media buzz can predict the box office success - more importantly based on the trending of the movie, strategies can be formulated to ensure favorable positioning of the movie. So in future, we can expand our analytics geographically to know how viewer's preferences vary with location. Moreover, we will add new constraints like retweets and so on to increase the efficiency of our analytics.

VII. CONCLUSION

We are still working on our conclusion since are analytics is not yet complete.

VIII. ACKNOWLEDGMENT

First and forempost, we would like to express our gratitude towards Professor Suzanne McIntosh for providing us with the opportunity to work on this project and also for her invaluable guidance and support during the course of this project. We would also like to express our very great appreciation to the Teaching Assistants of the course for helping us throughout the various sections of the project. We would also like to extend our thank for the guidance provided by past students who had taken this course during previous semesters. It helped us approach the course work and the project work in a better manner.

IX. REFERENCES