Hype Analysis Prediction

Sajan Bang
Student, New York University
New York, USA
sajanbang@gmail.com

Shreya Pandey
Student, New York University
New York, USA
Sp5333@nyu.edu

***Abstract*—**

In the real-world, hype is created when an event is announced. The world of internet has in many ways ability to often affect the success and failure of the event. Therefore, in our project we are analyzing the hype, and we are trying to predict effects of hype on an event's success or otherwise. We are taking a movie release as an event and collecting the data from twitter and reddit (comments and tweets) and analyzing it to predict the next day outcome. Analyzing the hype created by people can help in predicting the next day foot fall for the movie.

I. Introduction

Hype Prediction of movies is based on sentimental analysis of twitter and Reddit data from the time trailer is launched. These analysis can be used to predict the probability of movie to be hit, flop or average. We have used map-reduce for profiling and cleaning the data like removing special characters, retweets, etc. After this we used NLTK for removing the stop word so the prediction can be accurate and in the next process we used Textblob for sentimental analysis which will give the result per sentence as positive, neutral or negative with respect to date. Further we will calculate the scores of the positive, neutral and negative statements using Hive to find the overall result of the sentimental analysis in generating hype rating and actual rating. This will help in predicting the hype for the movie on the basis of reviews and tweets.

II. Motivation

Movies are the biggest source of entertainment and every year directors spend millions of dollars on movie advertisement and creating a hype in order to increase the probability of movie success. We will analyze if there is actually a relationship between hype and movies success. This will save a lot of money and help us understand genre of movies that people actually like and are hit at the box office. And also the pre rating based on hype can help the viewers in getting the insides about the movie's goodness.

III. Related Work

A Survey of Techniques for Sentiment Analysis in Movie Reviews and Deep Stochastic Recurrent Nets explore the task of sentiment analysis by using dataset provided by Socheretal (2013), which analyzes reviews on Rotten Tomatoes and includes 11,855 sentences and 215,154 unique phrases. A number of different techniques in analyzing the dataset is studied. As a baseline classifier, they implement Naive Bayes and will show that they can see substantial performance gains by encoding more information into their model. They explore both recursive networks that used the parsed tree structure, and build up sentiment predictions from that, as well as recurrent neural networks, which analyze each sentence and phrase as a varying length sequence with a single label.

MapReduce Functions to Analyze Sentiment Information from Social Big Data paper proposes a method to extract sentiment information from various types of unstructured social media text data from social networks by using a parallel Hadoop Distributed File System (HDFS) to save social multimedia data and using MapReduce functions for sentiment analysis. The proposed method has stably performed data gathering and data loading and maintained stable load balancing of memory and CPU resources during data processing by the HDFS system. The proposed MapReduce functions have effectively performed sentiment analysis in the experiments. Sentiment analysis is processed with the following steps using parallel HDFS and MapReduce functions. First, social networking big data is gathered from some SNS services. Second, the necessary data is extracted from the gathered data. Third, the extracted data is processed to load into the HDFS. Fourth, the processed data is loaded into the parallel HDFS. Fifth, sentiment analysis is processed via the MapReduce functions using dictionaries for sentiment analysis. Some experiments are processed for performance analysis of the proposed system and functions. The results of sentiment analysis with the proposed system are very close to the results of sentiment analysis with a manual process.

Sentiment Analysis of Twitter Data paper introduce two resources which are available 1) a hand annotated dictionary for emoticons that maps emoticons to their polarity and 2) an acronym dictionary collected from the web with English translations of over 5000 frequently used acronyms. They discuss classification tasks like sentiment analysis on micro-blog data. They give details about the data. they discuss their pre-processing technique and additional resources. They present their prior polarity scoring scheme. They present the design of their tree kernel. They give details of their feature-based approach. They presented their experiments and discuss the results. They presented results for sentiment analysis on Twitter. They used previously proposed state-of-the-art unigram model as our baseline and report an overall gain of over 4% for two classification tasks: a binary, positive versus negative and a 3-way positive versus negative versus neutral. They presented a comprehensive set of experiments for both these tasks on manually annotated data that is a random sample of stream of tweets.

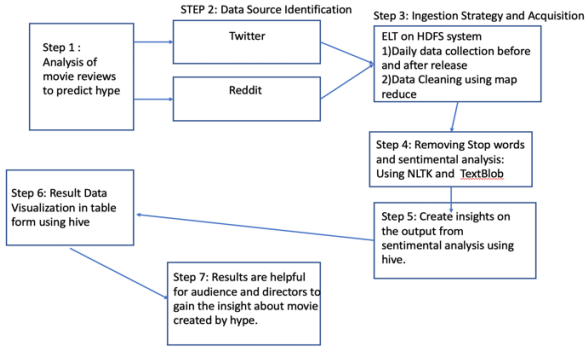
Twitter as a Corpus for Sentiment Analysis and Opinion Mining paper is pertinent to observe that with emergence of Web 2.0, the internet has evolved into a space where people are talking about everything, all discussions are a mine of data which lead to effective observations. The paper discusses rising demand of opinion analysis which is relevant to my project on analysis of opinions on Movie Reviews and making relevant predictions as per movie's current popularity and trends. The paper discusses certain implications of models to improve accuracy, it was considered that the winning strategy was to consider the last sentence of the document as sentiment. The corpus collection

was done by using regex for emoticons, tag distributions for pairwise comparisons. After that it was observed that objective texts contain more common and proper nouns whereas subjective texts were more often personal pronouns. Furthermore, the paper discusses methods indulged in identifying features for our dataset, and visualizations to observe improvement of accuracy overtime.

IV. Design and Implementation

A. Design Details

We have used waterfall model for developing this prediction code. Our source of data are Reddit and Twitter, we scrapped the data using API. Then we used map reduce method to clean and merge the data date wise which is a crucial step for the further predictions. We deliberately formatted the date in such a way so that we use them in easy way for the further step where we used NLTK and TextBlob for removing stop words and getting the sentiments for each sentence date wise. For the final analytics part we used Hive, where we worked on numbers to get final output as percentage, total count and rating on the basis of number of positive, negative and neutral sentiments for sentences.



B. Datasets

Two datasets are taken for this project one is from twitter and other is from reddit to get more accurate prediction. Twitter data is very unsorted and more into slang language. Cleaning and profiling play a vital role with twitter data as they contain emojis, special symbols, short statements. Whereas reddit data is more sorted as compared to twitter data. As this project provides prediction as per the dates i.e. daily, weekly and biweekly it becomes more important that schema should have two columns i.e. date and text (tweets and reviews). After cleaning & profiling, reviews must be merged as per dates to get the desired output.

V. Results

Initially when we started collecting data, we faced problems like the output files were of different format as the desired format. Our vision is to see how hype for a movie effects the foot fall for that movie and also how it will affect the box office collection for the movie. The scope of this project is not only limited to a particular movie, but we can also find the hype pattern and can predict the outcome of the movie even before its release. After performing our analysis, we got some good results with the error margin of +0.5 to -0.5. With this prediction tool we are able to calculate the hype rating before the movie release. We tried this application on a movie named as "venom". Our final results come out to be almost accurate with the +0.5-error margin.

VI. Future Work

Scope of this project is not limited, we will analyze multiple movies and try to predict the movie outcome. Also, we plan to find pattern and analyze different genre of movies to identify which movies will be popular among people.

VII. Conclusion

Through this analytics we will be predicting how hype effect rating is relating to the actual rating of the movie. Also, how footfall varies through hype with respect to hype rating. We will be using various online polls and box office collection data to check if the prediction though this approch is accurate(with +0.5 to -0.5 error margin) and good.

Acknowledgment

We are grateful to twitter and reddit for their contribution to this society where data plays a vital role and these social platforms are making it available to everyone. Apart from this we are also thankful to our Professor Suzanne Mcintosh who helped us with the insights of this project and gave us the future work we can do through this project. Also, we would like to thank NYU HPC for their technical contribution and keeping the cluster live all the time and sorting all the glitches with almost no time.

References

1. T. White. Hadoop: The Definitive Guide. O'Reilly Media Inc., Sebastopol, CA, May 2012.
2. A. Gates. Programming Pig. O'Reilly Media Inc.,Sebastopol, CA, October 2011.
3. J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. In proceedings of 6th Symposium on Operating Systems Design and Implemenation, 2004.
4. S. Ghemawat, H. Gobioff, S. T. Leung. The Google File System. In Proceedings of the nineteenth ACM Symposium on Operating Systems Principles – SOSP '03, 2003.
5. Alexander Pak, Patrick Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining
6. Chase Lochmiller, Department of Computer Science Stanford. A Survey of Techniques for Sentiment Analysis in Movie Reviews and Deep Stochastic Recurrent Nets
7. Ilkyu Ha, Bonghyun Back, Byoungchul Ahn. MapReduce Functions to Analyze Sentiment Information from Social Big Data, First Published June 1, 2015.
8. Apoorv Agarwal Boyi Xie Ilia Vovsha Owen Rambow Rebecca Passonneau ,Department of Computer Science Columbia University. Sentiment Analysis of Twitter Data