**TITLE :** Forecasting success of a movie using big data analytics and sentiment analysis.

**PROJECT AREA :** Big Data Analytics

**SPONSORED BY :** Persistent Systems Ltd

**GROUP MEMBERS:**

1. Vibha Ravichandran 9730 53 3079 ravichandranvibha@gmail.com
2. Ustat Kaur 9049 87 9039 ustat93@gmail.com
3. Tushar Prasad 8149 33 4983 tusshaarprasad@gmail.com
4. Sameer Thigale 9158 70 2389 sameerthigale08@gmail.com

**INTERNAL GUIDE:**

Prof. Reena Pagare

**COMPANY GUIDE:**

Mr. Chintamani Gokhale

**ACM KEYWORDS**:

Statistical databases, Forecasting, Big data analytics, Data mining, Sentimental Analysis..

**ABSTRACT:**

Social media content contains rich information about people’s preferences. Use of socially generated “big data" to access information about collective states of the minds in human societies has become a new paradigm in the emerging field of computational social science. A natural application of this would be the prediction of the society's reaction to a new product in the sense of popularity and adoption rate. However, bridging the gap between “real time monitoring" and “early predicting" remains a big challenge. An example is that people often share their thoughts about movies using Twitter. In this project we propose to do the data analysis on tweets about movies to predict several aspects of the movie popularity. The main results we present are whether a movie would be successful at the box office.

**CURRENT SCENARIO:**

Twitter, a micro blogging website, now plays an important role in the research of social network. People share their preferences on Twitter using free-format, limited-length texts,

and these texts (often called “tweets”) provide rich information for companies/institutes who want to know about whether people like a certain product, movie, or service.“Opinion mining” by analyzing the social media has become an alternative of doing user surveys, and promising results show that this could even be more effective than user surveys.

However, how to build an engine to detect and summarize user preference accurately remains a challenging problem.

**A BRIEF OUTLINE:**

In this work, we try to predict the movie popularity from sentiment analysis of Twitter data talking about movies. He we consider the tweets from the viewer. We manually label tweets to create a training set, and train a classifier to classify the tweets into: positive, negative, neutral, and irrelevant. We further develop a metric to capture the relationship between sentiment analysis and the box office results of movies. Finally we predict the Box Office results by classifying the movie as three categories: Hit, Flop, and Average. Our project also includes investigation on related topics like the relationship between tweet sent time and tweet number.

**PREVIOUS WORK IN THIS FIELD:**

The topic of using social media to predict the future becomes very popular in recent years. People have tried to show that twitter-based prediction of box office revenue performs better than market based prediction by analyzing various aspects of tweets sent during the movie release. They use twitter and YouTube data to predict the IMDB scores of movies. Sentiment analysis of twitter data is also a hot research topic in recent years. While sentiment analysis of documents has been studied for a long time, the techniques may not perform very well in twitter data because of the characteristics of tweets. The following are a few difficulties in processing twitter data: the tweets are usually short (up to 140 words). The text of the tweets is often ungrammatical. It investigates features of sentiment analysis on tweets data. However, few works directly uses sentiment analysis results to predict the future. People did sentiment analysis, but do not explicitly use the sentiment analysis results to predict the movie success.

**METHODOLOGY:**

We use sentiment analysis results of tweets sent during movie release to predict the box office success of the movie. Our methodology consists of four steps:

1. Data collection
2. Data Pre-Processing
3. Sentimental Analysis
4. Forecasting

**Data collection:**

We download an existing twitter data set and retrieves recent tweets via twitter API.

The following data fields of each tweet are stored:

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

The collected data is stored asa text file for each movie. The data fields are separated by tab.

**Data pre-processing:**

Since we have huge amount of data, we process them using distributed computing techniques. We further filter the data and get the tweets talking about movies via regular

expression matching. The goal of our data pre-processing consists of two major parts:

Part I, we need to get the information related to our prediction task.

Part II, we want toconvert the data to the format required by the input of our sentiment analysis tools (or extract the features required).

Data preprocessing is a great challenge in our task due to the data size.

We perform the filtering in three steps:

Step 1: we split the dataset into small chunks. Each chunk contains around 25,000 tweets.

Step 2: we assign basically the same number of chunks to every cluster node we use.

Step 3: For each node, we process the chunks and record the tweets related to target movies by regular expression matching.

Step 4: We combine the resulting tweets together.

The process is quite similar to Map Reduce framework and can be implemented also in Hadoop. We use HPCC cluster in our experiments, which already runs the portable batch system (PBS). Since we are able to ask for the nodes using PBS commands, there is no need to use Hadoop.

Another reason for not using Hadoop is that installing hadoop may require root access of the cluster which we don’t have. (This difficulty might be overcome using MyHadoop, which runs Hadoop above the PBS system) After we obtain the tweets related to 30 movies, we store them separately in 30 files. Then we sort them by the date the tweet was sent, and further get the tweets sent two weeks before and four weeksafter the release date of the movie

These tweets reflect the sentiment people have around the movie-release period. We use them in our prediction task.

We then delete the tweets which were not sent in English.

**Sentiment analysis:**

We train a classifier to classify tweets in the test set as positive, negative, neutral and irrelevant. We use Lingpipe sentiment analyzer to perform sentiment analysis on twitter data. The analyzer classifies the document by using a language model on character sequences. The implementation uses 8-gram language model. To create the training set and data for evaluation, we label the tweets based on the sentiment they carry. Following we have four categories:

Positive - Positive review of the movie

Negative -Negative review of the movie

Neutral-

* + - Neither positive nor negative reviews
    - Mixed positive and negative reviews
    - Unable to decide whether it contains positive or negative reviews
    - Simple factual statements
    - Questions with no strong emotions indicated
    - Hyperlinks / all external URL’s

Irrelevant –

* + - Not English language
    - Not on-topic (e.g. spam)

Since the number of tweets is huge and we lack enough human labor to manually label all of them, we randomly pick 200 tweets from each movie’s tweets sent in the critical period, and label them .We use simple word count as the feature. The accuracy is lower than Lingpipe so we do not use it in our experiment.

**Forecasting:**

We use the statistics of tweets’ labels to classify the movies as hit/flop/average.

Our prediction is based on the statistics of the tweets’ sentiment labels. We classify the movies as three categories: hit, flop, and average. We define hit as the circumstance that the profit of the movie is larger than its budget (>=20M), flop as the circumstance that the profit of the movie is less than its budget. Average case is 0<=Profit-budget<=20M. We develop a simple metric called PT-NT ratio to predict the movie categories of the success. According to the positive/negative/neutral/irrelevant tweets in the 200 randomly picked sample tweets, we can get the ratio of each category. We further use this ratio to estimate the total positive tweets, negative tweets, neutral tweets, and irrelevant tweets. We define the PT-NT ratio as total positive tweets/total negative tweets. Similarly, PT ratio is the percent of positive tweets, and NT ratio is the percent of negative tweets. We calculate the PT-NT ratio for each movie. We also calculate the profit ratio for each movie for comparison. The profit ratio = (revenue-budget)/budget. We use a hard threshold to determine a movie’s success.

* PT-NT Ratio (more than or equal to 5): Movie is hit
* PT-NT Ratio (less than 5 but more than 1.5): Movie would do Average business
* PT-NT Ratio (less than 1.5): Movie is Flop

Although this metric and simple and preliminary, it corresponds well with the real movie categories in our experiments.

**EXPERIMENTS:**

Our experiment consists of three parts:

1. We investigate the relationship between tweet number and the sent time, and show that the tweets number about the movie clearly reaches its peak around the movie release.

2. We show that the PT-NT ratio curve has the same tendency as the profit ratio.

3. We show the sentiment analysis results using Lingpipe.

4. We predict movies released and evaluate our prediction by statistics till date

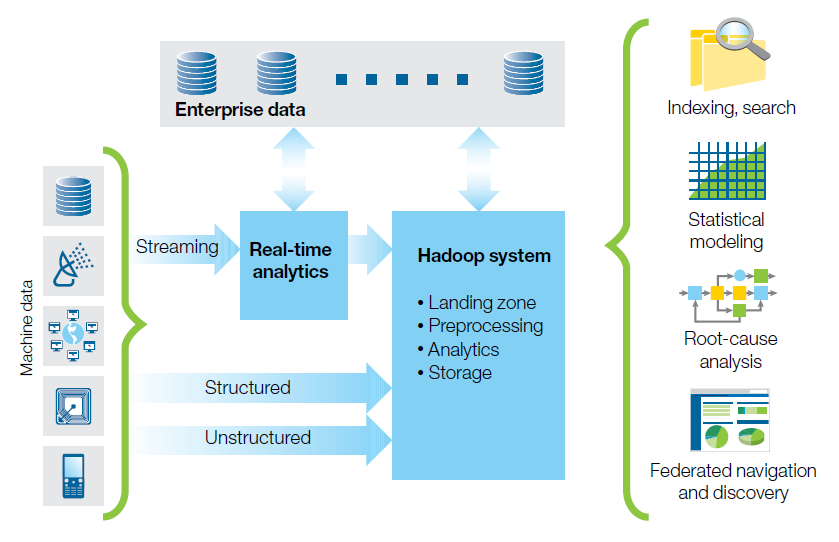
** **

Figure 1:Tweets Number Figure 2: Graph Showing Movies Vs P/N Ratio

**PROPOSED ARCHITECTURE:**

Since we are primarily dealing with big data, our software architecture will be based on the data set held by us, the fields data set contains and the other properties of the data set. **Hadoop** is the excellent platform for Big data analytics. The data set will be organised using the **Hadoop Distributed File System (HDFS).** Software architecture may primarily be a Hadoop Cluster running on top of Ubuntu OS (Since open source). We also propose to use **Apache Hive** for firing SQL queries or plan to use Cloud compute engines like **Google** **Cloud Compute** or **Amazon Web Services (AWS)** for performing computations on the go.

The end result will be a **WebApp** providing a UI for performing operations and firing queries.

****

**Figure 3: Operations analysis combines machine and enterprise data for rich insights.**

**PLATFORM / TECHNOLOGY:**

Software:

1. Apache Hadoop 2.5.0
2. Apache Hive 0.13.1
3. Ubuntu 14.04 LTS 64-bit
4. OpenJDK 7
5. Mozilla Firefox 30.0
6. LingPipe

Hardware:

1. Commodity hardware
2. Hardware for forming a computer network

**BENEFITS AND APPLICATIONS:**

* Forecasting rich insights about the success of the movie
* Application will help generate the sentimental analysis of the viewer.
* Help organise viewer response and reaction data and Meta data.

**DRAWBACKS OF THE SYSTEM:**

* Large processing time
* There are limitations of Twitter APIs (e.g. 1500 tweets/day). We do not have enough computing resources to crawl the data, which might result in an inaccurate number of tweets.
* Lot of spam and noise included in randomly picked 200 tweets.
* We do not take the total tweets number into account in our prediction metric.
* The sentiment analyzers we use have rather low accuracy.

**REFERENCES:**

[1] Prediction of Movie Success using Sentiment Analysis of Tweets ,Vasu Jain, Department of Computer Science,University of Southern California,Los Angeles, CA, 90007 ,vasujain@usc.edu . The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE’13],San Francisco State University, CA, U.S.A., March 2013

[2] Forecasting consumer behavior with innovative value proposition for organizations using Big Data Analytics, *A Balar, N Malviya, S Prasad, A Gangurde, IEEE 2013*

[3] An Oracle White Paper December 2013, Utilities and Big Data: Using Analytics for Increased Customer Satisfaction

[4] Early Prediction of Movie Box Once Success based on Wikipedia Activity Big Data Marton Mestyan1, Taha Yasseri1;2;3;\_, Janos Kert\_esz1;3;4,Institute of Physics, Budapest University of Technology and Economics, Budapest, Hungary, Oxford Internet Institute, University of Oxford, Oxford, UK, Department of Biomedical Engineering and Computational Science, Aalto University,

Aalto, Finland Centre for Network Science, Central European University, Budapest, Hungary, 26th June 2013.

[5] FORECASTING-Methods and Applications by- Spyros Makridakis,Steven C.Wheelwright,Rob J. Hyndman