The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.46 e-ISSN: 2251-7545

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

***Abstract*—Social media content contains rich information about people's preferences. An example is that people often share their thoughts about movies using Twitter. We did data analysis ontweets 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.**

I. INTRODUCTION

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.

In this work, we try to predict the movie popularity from sentiment analysis of Twitter data talking about movies. We analysis both the tweets in 2009 and recent tweets in 2012. 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.

II. RELATED WORK

The topic of using social media to predict the future becomes very popular in recent years. [1]tryto 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. [2] uses 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. [6] investigates features of sentiment analysis on tweets data.

However, few works directly uses sentiment analysis results to predict the future. [1] did sentiment analysis, but do not explicitly use the sentiment analysis results to predict the movie success.

III. 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:

A. Data collection

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

We download the 2009 dataset via a link (now expired) provided by SNAP research group at Stanford University (http://snap.stanford.edu). For the 2012 dataset,we use a python library Tweepy[7] which provides the access to Twitter API to retrieve data from Twitter.

We use the streaming API of Tweepy to get the tweets relevant to our task. The API, tweepy.streaming.Stream, continually retrieves data relevant to some topics from Twitter's global stream of Tweets data. The topics we use are a list of keywords related to the movie, e.g. "skyfall" and "wreckit Ralph". We list the keywords in Section 4.

The following data fields of each tweet are stored:

- Tweet Id
- Username of person who tweeted
- Tweet text
- Time of tweet

Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors*).

The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.46 e-ISSN: 2251-7545

- 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.

B. 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 preprocessing consists of twomajor 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. In the 2009 dataset, we deal with around 60GB of raw data. In the 2012 dataset, we have around 1GB of raw data. So it's important that we use big-data analysis techniques.

For the 2009 dataset, we need to first filter the datasets and get the tweets related to our target movies. Due to the large data size, we run the job in parallel on multiple nodes of the cluster.

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 MapReduce[8] 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.(We will refer to this period as "critical period" in Section 4) 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. We use the language detection tool at https://github.com/shuyo/ldig.

"Noisy" tweets are unavoidable in our data. They are tweets which contains the movie keyword but have nothing to do with the movie. For example, the following tweet is a noisy tweet for "Avatar":

#o2fail - grab your twitter backgrounds and avatars http://bit.ly/6e9Xa here. Show your support - 5000 signatures O2 still not moving.

We try to filter the noisy tweets by removing duplicates, since ads usually have very large amount of retweets. However, it is impossible to remove all noisy tweets automatically. In practice, we can remove them during the manual labeling process.

For the 2012 dataset, we can omit the filtering step because we already filter it during the data collecting process. However, we still need to perform other data preprocessing steps.

According to our prediction task, we need to get the tweets during the movie release. We define the "critical period" of the movie as the period between two weeks before the release date of the movie and four weeks after the release date. We sort the tweets according their sent time, and get the tweets sent in the critical period for our sentiment analysis task.

C. Sentiment analysis

We train a classifier to classify tweets in the test set as positive, negative, neutral and irrelevant.
We use Lingpipe sentiment analyzer [3] 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 [5], we have four categories: positive/negative/neutral/irrelevant. The labeling standards are as follows:

Table1 .

Positive	• Positive review of the movie
Negative	• Negative review of the movie
Neutral	• Neither positive nor negative reviews • Mixed positive and negative reviews

The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.46 e-ISSN: 2251-7545

	• 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. In the 2009 dataset, we randomly choose 24 movies (8 hit, 8 flop, 8 average) as the training set (4800 tweets in total). The other 6 movies are used as the test set. All 2012 data (8 movies, 200 tweets each) are used as another test set.

We train the sentiment analyzer using our training set and test on both 2009 and 2012 test sets.

We also train a naïve bayes classifier using NLTK toolkit [4] and the same training and test set. We use simple word count as the feature. The accuracy is lower than Lingpipe so we do not use it in our experiment.

D. Prediction: 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.

IV. 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 8 movies released in Nov, 2012 and evaluate our prediction by statistics till date

A. Tweets number vs. Tweet sent time

We investigate the ratio of "critical period tweets". It is computed by the number of tweets sent in critical period / the total number of tweets we have. (To save space, we do not include the detailed table) For most cases (20 out of 30 movies), this ratio is more than 50%. We investigate the exceptions and there are three main reasons:

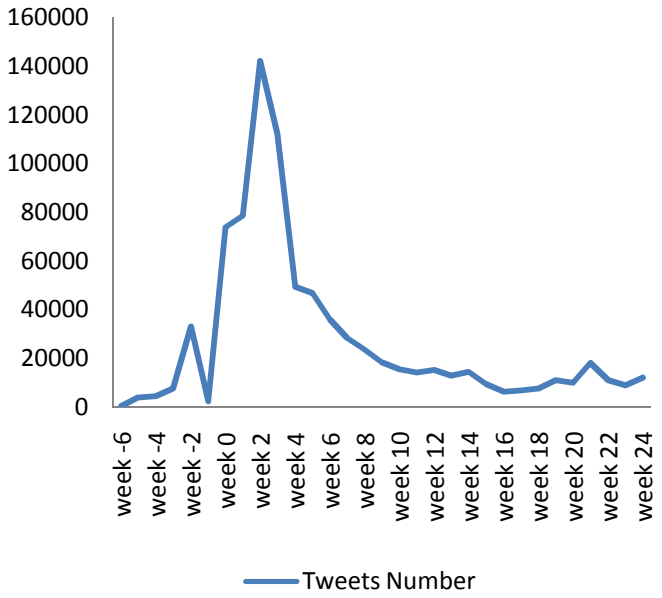
1) The movie name consists of very common words so there are a large number of "noisy" tweets. We already try to avoid movies like "Up", but it seems that the movie "Year One" still suffers from this problem.
2) The movie was released in June or December, so we lack some of the tweets sent in the critical period.
3) Some movies are really popular that people talk about them even they have been released for more than one month. For example, "Ice Age: Dawn of the Dinosaurs" and "The Twilight Saga: New Moon".

On average, "hit" movies receive more tweets than "flop" and "average" movies. This also coincides with our intuition.

We show the relationship between sent time and number of tweets on the "Harry Potter and the Half-Blood Prince" here. In the graph, the X axis is the week number, and the Y axis is the number of tweets sent during that week. We define week 0 as the week the movie was released. Week -2 means the two weeks period before the week 0, and week 2 means the two weeks period after the week 0, .etc. Due to space limits we use two-week period.

The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3n3.46 e-ISSN: 2251-7545

Fig.1

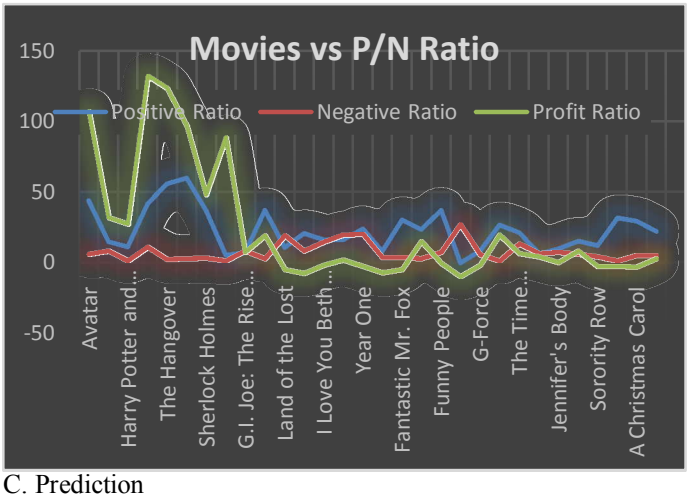


The figure coincides with our intuition very well. The number of tweets about the movie begins to increase as the movie releases, and gradually decreases after one or two months.

B. PT-NT ratio vs. Profit ratio

We plot the positive ratio, negative ratio, and profit ratio of the 30 movies released in 2009 in the following graph. We can clearly see some spikes in here which are in accordance with the success of a movie.

Fig.2



C. Prediction

We predict 8 movies which just released using PT/NT ratio, and the results are shown in the table.

Table2

Movie name	Positive	Negative	Neutral	Irrelevant	Correct	Correct %
G.I. Joe The Rise of Cobra	8	9	93	140		
	3	1	24	100	128/200	64%
Alvin and the Chipmunks The Squeakquel	53	4	86	57		
	26	0	77	27	130/200	65%
Funny People	42	8	63	87		
	25	1	41	55	122/200	61%
Ghosted	0	10	27	163		
	0	2	21	104	127/200	63.5%
A Christmas Carol	34	6	75	85		
	18	0	60	46	124/200	62%
Love Happens	23	5	75	97		
	9	1	60	71	141/200	70.5%
Average						64.4%

The first line of each movie is the number of tweets in each category (manually labeled), and the second line is the correctly predicted number of tweets in each category. We use the accuracy of all 200 tweets.

The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.46 e-ISSN: 2251-7545

The accuracy is 64.4%, which is better than [6](which results in 60.5%). However, our test set is different, so the comparison is not fair. Lingpipe sentiment analyzer is originally created for sentiment analysis of documents, so it is not the perfect choice for sentiment analysis on twitter data

C. Prediction

We predict 8 movies which just released using PT/NT ratio, and the results are shown in the table.

Table3

Movie	PT/NT Ratio	Prediction	Budget	Sales	Days since Release	Prediction Confidence
Wreck-it Ralph	24.5	Hit	$120 M	$158 M	32	Yes
Skyfall	4.08	Hit	$200 M	$246 M	25	Yes
Twilight Saga: BD II	47.53	Super hit	$120 M	$255 M	18	Yes
Rise of the Guardians	30.89	Hit	$145 M	$49 M	18	May be
Red Dawn	17.7	Hit	$65 M	$32 M	13	Yes
Miami Connection	20	Hit	$1 M	**	25	Can't say
Citadel	#####	N/A	N/A	***	25	Can't say
Nature calls	1.5	Avg	N/A	N/A	25	Can't say

We predicted 5 movies to be hit and one to be super hit, one to be average and we could not determine success rate for one due to it data unavailability. Comparing our prediction results with box office results till date we found our prediction to be exact for four cases, for a case it is on border line between hit and average and for one we could not find data to check our prediction confidence.

The results for classification of our test data using Lingpipe toolkit shows accuracy for labelling to be 64.4% which is within 15% of the accuracy for Lingpipe for movie review analysis. Movie review being a text with multiple sentences and greater length compared to maximum 140 character tweets, the classification algorithm might not be perfect for tweets.

D. Conclusion

We did some preliminary study in using sentiment analysis to predict a movie's box office success. The results show that the box office success can be predicted by analyzing sentiment of the movies with simple metrics and pretty good accuracy.

We understand that there might be more than one factor which affect the movie box office success, but we concentrate on sentiment analysis in this work. As sentiment analysis on twitter itself is a challenging topic, 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:

A. 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.
B. Lot of spam and noise included in randomly picked 200 tweets.
C. We do not take the total tweets number into account in our prediction metric.
D. The sentiment analyzer we use have rather low accuracy.

Future Work which may be done to improve reliability and accuracy in our prediction model:

A. Temporal analysis will be added in the project.

Use different other models and algorithms

REFERENCES

[1] Sitaram Asur&Bernardo A. Huberman. (2010) Predicting the Future with Social Media. Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01, pp. 492-499.

[2] Andrei Oghina, Mathias Breuss, Manos Tsagkias&Maarten de Rijke. (2012) Predicting IMDB movie ratings using social media. Proceedingsof the 34th European conference on Advances in Information Retrieval, pp. 503-507.

[3] Lingpipe sentiment analyzer:http://alias-i.com/lingpipe/demos/tutorial/sentiment/read-me.html

[4] NLTK: http://nltk.org/

The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.46 e-ISSN: 2251-7545

[5] Twitter Sentiment Corpus http://www.sananalytics.com/lab/twitter-sentiment/

[6] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau (2011). Sentiment Analysis of Twitter Data.Proceedings of the ACL 2011 Workshop on Languages in Social Media, pp.30-38

[7] Tweepy libray. http://tweepy.github.com/

[8] DEAN, J., AND GHEMAWAT, S. Mapreduce: simplified data processing on large clusters. In OSDI'04: Proceedings ofthe 6th conference on Symposium on Opearting Systems Design & Implementation (Berkeley, CA, USA, 2004), USENIXAssociation.