Prediction Of Movie Success using Sentiment

Analysis Of Tweets

**Aishwarya SS, Sweety Kumari, Raksha Singh, Priya Sharma**

Student, Computer Science, M S Ramaiah Institute Of Technology, Bangalore, India

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 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.

Keywords: Sentiment Analysis, Text Mining, Big Data, R programming

1. **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.

R studio is one of the many programmes that offer packages that can analysis data, however R studio works well with statistical problems and has a user friendly interface. The classification model which this project will develop will determine whether the tweet status updates (which cannot exceed 140 characters) reflects positive opinion or negative opinion on the behalf of the person who tweeted. This paper will use a hybrid of knowledge based sentiment analysis methodologies which have been more traditionally used, and those of R programming methodologies which used a more intuitive approach to sentiment. The results of these two methodologies will be used to perform a thorough analysis of the dataset.

1. **DESIGN**

This paper proposes a hybrid approach involving both knowledge based methodologies and machine learning based methodologies to analysis the sentiment orientation of the tweets. Tweets are accessed through a Twitter API. The collected reviews are proposed in R and saved to an excel file. The words were extracted and stored in a feature vector. The words were scored into their relative sentiment orientation using a sentiment lexicon which was sourced from the internet and was publicly available. The excel file was then entered into R and a corpus was produced. The corpus was then divided into a ‘training set’ and a ‘test set’ and feature were extracted from each respectively. From the ‘training set- feature extraction’ a machine learning algorithm was produced from the ‘test set-feature extraction’ a model classifier in the form of Naive Bayes classifier was produced. Naive Bayes does not consider the relationships between features such as emotional keywords and emoticons. This is ideal for sentiment analysis as often these features do not always relate to one another such as in the use of a smiley emoticon at the end of a negative tweet. Naive Bayes Classifier analyzes each of the features of the feature vector individually as it assumes that they are equally independent of each other. The conditional probability for Naive Bayes can be defined as

P (X |yj ) = Πm P (xi |yj

’X’ is the feature vector defined as X={x1 ,x2 ,....xm } and yj is the class label. In the tweets collected for this paper there are different independent features such as emoticons, emotional keyword, which are treated as either positive or negative and so are utilized by Naive Bayes classifier for classification.

1. **IMPLEMENTATION**

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

We download an existing twitter data set and retrieves recent tweets via twitter API.We use the streaming API of Twitter to get the tweets relevant to our task. The API, tweepy .streaming .Stream ,continually retrieves data relevant to some topics fromTwitter’s global stream of Tweets data. The topics we useare 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

1. 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 two major parts:

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

Part II, we want to convert the data to the format required

by the input of our sentiment analysis tools (or extract the

features required).

“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.

1. 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/negative/neutral/irrelevant. The labeling standards

are as follows:

TABLE 1

|  |  |
| --- | --- |
| Positive | * Positive review of movie |
|
|
| Negetive | * Negetive review of movie |
| Neutral | * No review of movie * Mixed review of movie |

1. Predection

Our prediction is based on the statistics of the tweets’

sentiment labels. We classify the movies as three

categories: hit, flop, and averageWe 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 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.

1. **SCOPE AND FUTURE WORK**

We proposed an approach which considers subjective sentences and objective sentences both but here we are applying this approach on close domain named “Movie Reviews”. We are expecting good efficiency for our proposed approach. In future we will apply this approach for the other close domain. We need a global SentiWord dictionary which we can use for either for open or close domain and contain proper threshold values. In future we can use the pure semantic, ontology and Description logic approach to classify subjective sentences and objective sentences and then after classify into positive, negative or neutral category.

VII.**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.

**ACKNOWLEDGMENT**

We would like to thank my guide Prof & Head Dr. K.G.Srinavasa for his valuable suggestions and encouragement. Special thanks to Asst Prof. Chandrika Prasad and Asst. Prof. Sawmya B J for giving me constant guidance and suggestions.

**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.Proceedings of the 34th European conference On Advances

in Information Retrieval, pp. 503-507

[3] Lingpipe sentiment analyzer:http://aliasi.

com/lingpipe/demos/tutorial/sentiment/read-me.html

[4] Twitter Sentiment Corpus

<http://www.sananalytics.com/lab/twittersentiment/>

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