Public Opinion Mining On Twitter

Jiangnan Li
PhD Student, Department of Electrical and Computer Engineering, UTK

Abstract

Twitter is one of the most popular microblogging applications in which user post messages, called tweet. Due to its huge amount of users and world wide impact on the Internet, Twitter has become an important platform to trend popular topics and an accurate indicator for Internet public opinion. Therefore, it is necessary to mining valuable information on Twitter.

To implement this, this project selected the 89th Academy Award as the objective event, collecting and processing related data from Twitter, and making prediction through data mining tools. The comparison between prediction and corresponding fact shows that public opinion on Twitter can reflect real event to an extent.

Introduction

With the rapid development of Internet, a number of social medias grew and have become an important part of people's social life. People like to post their thoughts, record their daily life events, and write comments in respect to special event or product on social media platforms. In addition, people are also likely to review other's comments on a special entity as reference before making a decision. These comments can be important in several ways for different group of communities.

Different from traditional survey which requires participant to answer a series of questions, opinions posed by users on their own initiative can provide more accurate feedback and real reflection for a specific entity. However, these data is always distributed over a huge remote database or even the entire Internet. Moreover, social media data always has a wide range, making it different to extract valuable information.

In order to analyze the public opinion on Twitter and its relationship with real fact, this project selected the 89th Academy Award as the target event, collecting and processing related tweets posted by Twitter users. Meanwhile, a text classifier was trained to implement sentiment analysis of each tweet. Finally, the project compared the prediction result with real facts.

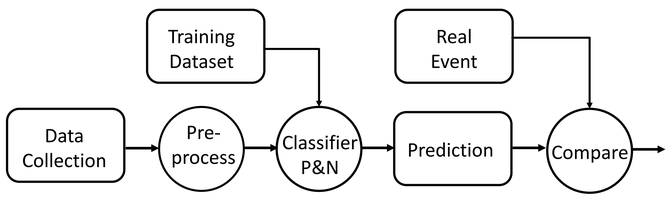

Figure 1. Project Overview.

Design and Implementation

This project firstly obtains a list of all followers of the Academy Awards official Twitter account through Twitter API. Then, each follower's tweets posted over a period of time were collected. These tweets are filtered and classified according to nominees of the twelve awards. After that, necessary pre-process procedures are implemented for each tweet to increase the classification accuracy, including tokenization, stop words, and repetition.

Meanwhile, a Naive Bayes classifier is trained for classification. This project uses a public dataset that contains tweets and corresponding sentiments to train a classifier. After the training process, a test dataset is used to test the accuracy of the classifier. Then, all the processed tweets are classified and tagged into positive and negative using the trained classifier.

By sum all the score together, we can get an overall score for each nominee.

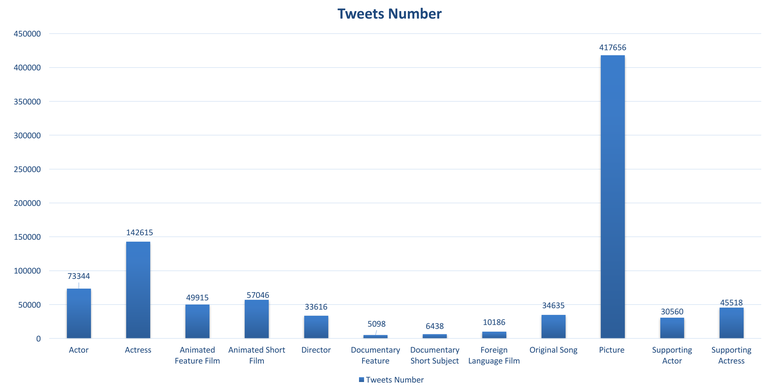

Figure 2. Tweets Number

Results

Totally 2433939 followers' information was obtained, and over 120 million tweets of these followers were collected. These tweets were filtered and classified according to each nominee of each awards. **After filtering, totally 906624 tweets which contain key words were left.**

The classifier training process using a data set contains over 200000 sentiment records. **The overall accuracy for the classifier was about 70% over different test data sets.** This project tests and tags each tweet of each nominee of each award, and sum up the score that each nominee earned, and finally, compares the nominees who earns the highest score with the real winner of corresponding awards. **According to the result, 7 out of 12 predictions are same to the real facts**.

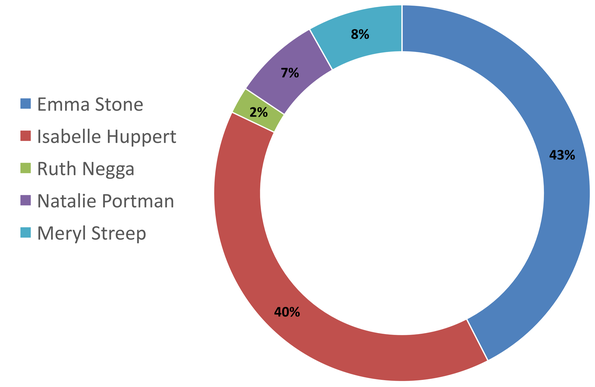

Figure 3. Prediction of the Best Actress Award

Discussion

The prediction seems not accuracy enough compared with real facts due to following reasons.

First, the tweets of followers are filtered and classified according to a list of key words which may contain ambiguity. Then, the accuracy of the Naive Bayes classification was not high enough. Finally, the classification showed that the scores of some nominees were really close, a tiny advantage was not persuasive compared with the accuracy of the classifier.

In further work, a larger data set should be used to train the classifier. Meanwhile, stemming and N-grams can also be added into the tweet pre-processing part. In addition, the tweets can be further filtered to remove tweets which do not contain sentiment by using related Natural Language Processing algorithms, so that only tweets contain specific grammar relationships will left. These methods will definitely increase the accuracy of classification to an extent.

Conclusion

This project analyzed the correlation between public opinion on Twitter and real event. The project collected related tweets posted by Twitter users through a web crawler, meanwhile, a Naive Bayes classifier was trained using a public sentiment training data set. The tweets were then classified and tagged by the Naive Bayes classifier. An overall opinion of tweets was obtained by summing up all tags. Experiment result showed that the accuracy of the classifier is around 70% and 7 out of 12 prediction results are same as real facts. This indicates that the public opinion on Twitter can reflect real event to an extent.

Contact

Jiangnan Li
PhD student, EECS Department, UTK
jli103@vols.utk.edu

References

1. Hridoy, Syed Akib Anwar, et al. "Localized twitter opinion mining using sentiment analysis." *Decision Analytics* 2.1 (2015): 8.
2. Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." *CS224N Project Report, Stanford* 1.12 (2009).
3. Lee, Kathy, et al. "Twitter trending topic classification." *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*. IEEE, 2011.
4. Jiang, Long, et al. "Target-dependent twitter sentiment classification." *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011.
5. Gokulakrishnan, Balakrishnan, et al. "Opinion mining and sentiment analysis on a twitter data stream." *Advances in ICT for emerging regions (ICTer), 2012 International Conference on*. IEEE, 2012.
6. Roesslein, Joshua. "tweepy Documentation." *Online] http://tweepy. readthedocs. io/en/v3* 5 (2009).
7. Bird, Steven. "NLTK: the natural language toolkit." *Proceedings of the COLING/ACL on Interactive presentation sessions*. Association for Computational Linguistics, 2006.