Sentiment Analysis of Real-Time Event with Twitter

Kai Qi,He Huang,Yiying Zhang,Yajing Wu

1. Problem Statement

Individuals make decisions greatly influenced by others. These influentials can be celebrities, friends or colleagues. If lots of people talk about one movie, individual will be aware of it. If colleagues or friends watched and talked about one movie, individual tend to buy a ticket for that movie. Twitter is a natural environment to study this phenomenon which is also called word of mouth(WOM)[1]. In this project, we want to investigate how twitter moods influence movie office sales in the real world. To achieve this goal, we have two main tasks: 1) classify the sentiment of tweets accurately 2) study the relationship between twitter mood and office sales of movies.

2. List of possible approaches with citations to relevant work

In this project, we use data from Sentiment140[2] to train and evaluate classifiers. It contains 1.6 million training tweets and 500 test tweets. The training data was automatically collected from twitter. They assume that tweets with symbols like ':)' are positive and tweets with ':(' are negative. And test data is manually labeled.

Twitter streaming API offer functions to filter real time data with keywords. However, collected tweets may be not what we want. For example, there is one movie called 'Focus' released on 27 Feb 2015. We filter the streaming data with keyword 'Focus' and get lots of unrelated tweets. Thus, each tweet needs to be decide whether relevant or not. In previous work[3], to determine when a tweet was an advertising tweet, they simply checked whether the tweet contains a URL. And to reduce irrelevant tweets, they used an two step approach. They have an movie dictionary with related phrases like 'movie', 'cinema'. The first step is to pick out tweets containing phrases in dictionaries. And for each movie they created a customized dictionary to eliminate irrelevant tweets. For example, for movie 'Focus', if one tweet contains the phrase 'focus on', that tweet tend to be not relevant. In [4], with streaming data in twitter, they designed a system to detect earthquake in real time. They trained a classifier(SVM) to decide whether the tweet is relevant or not. For each tweet, they prepared three kinds of features: A) the number of words in tweet B) keyword C) words before and after keywords. In [5], a naive bayesian classifier is used to filter out irrelevant tweets.

After eliminating advertise and irrelevant tweets, we need to distinguish between intention tweets and other tweets. Intention tweets show that authors have not seen that movie but want to watch. In [5], a two step approach is applied. In first step, intention tweets are classified by a dictionary for intention words and non-intention tweets are classified by a dictionary for non-intention words. In second step, the rest of tweets which can not be classified by dictionaries are classified by a bayesian classifier. In [3], intention tweets are classified by a SVM classifier. This classifier was trained and tested on a dataset of more than 3000 manual labeled tweets. At last, the rest of tweets are classified into positive, negative and neutral. In the step above, a naive bayesian classifier is used in [3][4]. The processes of tweet classification in [3] and [5] are similar, only differ in some details.

In this project, we started to collect tweets with keywords 'Focus', 'the lazarus effect' and 'outcast' from Feb 26 00:00:00 and these movies were released on Feb 27. By checking the

content of tweets, we found that many tweets which contained URL were not advertising tweets. After discussion, we decided to filter out irrelevant tweets with keywords. To classify advertising tweets and intention tweets, a multinomial naive bayes classifier will be used.

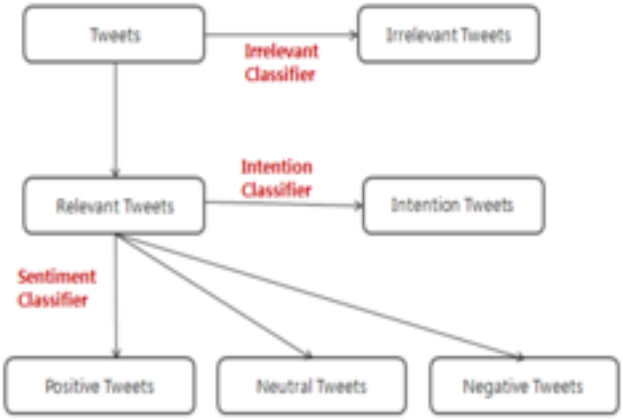

Fig 1: the process of tweet classification, come from[5]

3. Project plan for the rest of the term

For next a few weeks, we will continue to collect tweets with keywords ‘Focus’ and ‘the lazarus effect’. And at the same time, daily movie sales data will be collected from BoxOfficeMojo.com. To classify advertising tweets and intention tweets, a dataset will be built and around 1000 tweets will be manually labeled.

To get a better result, we will compare the performance of three different classifiers and evaluate classifier with different metrics. Three classifiers are support vector machine, multinomial naive bayes and k nearest neighbour classifier. A discussion of different classifiers will be included in the final report. In paper [3], a dynamic panel data model is used to analyze the relationship between movie sales and twitter moods. They found that positive Twitter evaluation is associated with higher movie sales. In paper [5], LSDV analysis is applied to analyze impact of tweets on movie revenue. Weekly tweet volume trend is also studied. And they found that movie revenue was higher when the volume of tweets was larger during before and after film release. We will learn and apply one of these methods in this project. And a discussion of analysis result will be included in final project.

4. References

[1] E. Katz, P. Lazarsfeld, Personal Influence, Free Press, 1955
[2] A. Go, R. Bhayani, L. Huang, sentiment140 corpus[Data set], 2009
[3] Rui, Huaxia, Yizao Liu, and Andrew Whinston. "Whose and what chatter matters? The effect of tweets on movie sales." Decision Support Systems 55.4 (2013): 863-870.
[4] Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. "Earthquake shakes Twitter users: real-time event detection by social sensors." Proceedings of the 19th international conference on World wide web. ACM, 2010.
[5] Baek, H. M., JoongHo Ahn, and S. W. Oh. "Impact of Tweets on Movie Sales: Focusing on the Time when Tweets are Written." Journal of ETRI (2014).