Sentiment Analysis of Real-Time Event with Twitter

Kai Qi,He Huang,Yiying Zhang,Yajing Wu
NLP CSCI 4152/6509

1. Abstract

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 tends 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 investigated how twitter moods influence movie office sales in the real world. To achieve this goal, movie tweets are collected from twitter streaming API and irrelevant tweets are filtered out. The rest tweets are classified into four categories (intention, positive, negative and neutral) and then panel data analysis is performed. At last, we found that positive tweets have good influence on movie sales and negative tweets have bad influence. Another interesting found is that tweets, whose author express intention to watch one movie, have the greatest influence on movie sales.

2. Introduction

Influence has been studied in areas like sociology, marketing for a long time. It can help companies and politicians design better campaign strategies or help researchers better understand how society works.

Many works have been done to understand the diffusion of influence. In a traditional theory, some individuals, also called influentials[2], are good at spreading their opinions and have influence on others' decision. With these influentials' help, companies may better sell their products. Celebrity endorsement is an example. And in a more modern view, some researchers state that people tend to make decisions based on the opinions from their friends or peers and they believe "network-based advertising"[2] is a better choice. Word of mouth (WOM) is an example of influence diffusion.

Things became a little different when it comes to new age. Individuals are more likely to get influenced by others since it is more convenient for them to communicate with others and it appears more often and in a way that is more direct. Social networks like Twitter or Facebook offer an easy way for individuals to update statues and receive statues from friends, colleagues or idols and many social activities become online. Take Twitter as an example, it has over 250 million active users and about 700 millions new tweets per day, which offer researchers an information rich environment to study word-of-mouth(WOM).

In this project, we investigated the relationship between twitter mood and movie sales. Three movies' tweets are collected from Feb. 27 to April 4 and daily movie sales are

collected from BoxOfficeMojo.com. To classify advertising and intention tweets, a tweet sentiment dataset is built and around 3000 tweets are manually labeled. And two additional datasets are used for sentiment analysis. Three different classifiers are used and compared. At last, panel data analysis is applied to study the impact of tweets on movie sales.

3. Related Work

Many studies on the impact of tweets have been done. Some work focus on twitter mood and study its influence on events in real world. [3] investigates whether twitter mood are correlated to the values of stock market. In [4], sentiment analysis was applied to political tweets and results were used as a predictor of the election result. Some works focus on relationship among users and tweets. In [2], with different measures, authors studied user influence on different topics.

Early research on stock market prediction was based on random walk and efficient market hypothesis. According to efficient market hypothesis, stock market prices are largely driven by real time news. Since news is unpredictable, stock market prices will follow a random walk pattern and cannot be predicted. However, behavior economics states that financial decisions are also driven by emotional moods. Based on this theory, researchers studied how public moods influence the stock markets [3]. Sentiment of tweets is analyzed and organized as daily time series of public mood. Since the relation between public moods and the stock market is almost non-linear. A neural network is trained to make predictions. And mood time series and historical data of DJIA were used as input. At last, they got an accuracy of about 87% in predicting daily up and down in the closing value of the DJIA.

Word-of-mouth (WOM) is often considered as credible information to consumers. In [1], authors studied whether and how twitter mood influence movie sales with a panel data model. They found that positive WOM is always related to good movie sales and negative WOM is related to bad movie sales. Another result is that some tweets, of which the author express intention to watch a movie, have the greatest influence on movie sales. [5] studied the same problem in a slightly different way and got similar results.

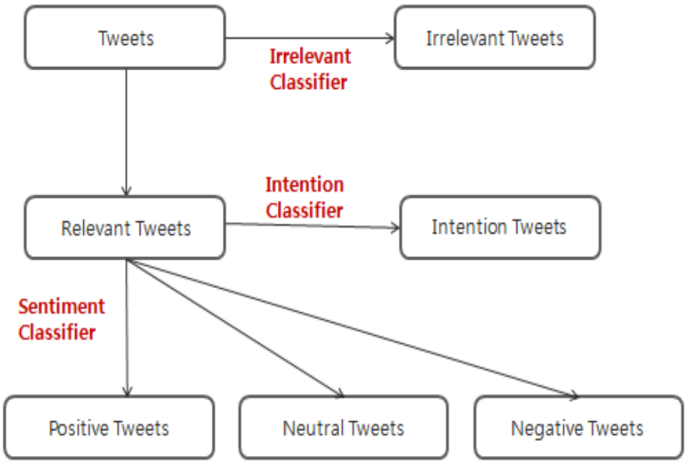


Fig 1: the process of tweets classification[5]

The processes of tweet classification in [1] and [5] are similar, only differ in specific methods. In [1], to determine whether a tweet is an advertising tweet, they simply checked whether the tweet contains a URL. A two-step approach was proposed to eliminate irrelevant tweets. A movie dictionary was built, which contains phrases like “movie”, “cinema”. The first step is to pick out tweets containing phrases in the dictionary. 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 tends to be irrelevant. Intention tweets show that authors have not seen that movie but wish to watch. In [5], a two-step approach is applied to classify intention tweets. In first step, intention tweets are classified by a customized dictionary. In second step, the rest of tweets are classified by a naive bayesian classifier. In [1], intention tweets are classified by a SVM classifier. This classifier was trained and tested on a dataset contains more than 3000 manual labeled tweets. At last, non-intention tweets are classified into positive, negative and neutral.

In [6], authors studied the real-time events in Twitter and proposed an algorithm to detect a target event like earthquake. A system is built to detect an earthquake or typhoon occurrence and it sends warning emails to related Twitter users. When earthquakes happened, users got warned in less than 2 minutes, which is much faster than professional organizations like Japan Meteorological Agency (JMA). In this work, each twitter user is considered as one sensor and tweets from one user are considered as sensory information. And then the detection problem turn into object detection and location estimation. To obtain tweets on target events, a support vector machine is used to decide whether tweets are related. And then a probabilistic model is built to the centre and trajectory of the event location.

4. Problem Definition and Methodology

In this project, tweets are classified into four categories (intention, positive, negative and neutral) and then a panel data model is used to analyze the relationship between movie sales and twitter mood. Most of work in this project is based on [1] and [5] but several improvements have been made. Instead of using weekly movie sales, daily movie sales and daily tweets have been studied here. Thus there are more observations and WOM is studied in a smaller time granularity. Instead of using historical data, we studied recent movies and collected tweets in real-time. Thus we can better understand individuals’ thoughts at present and have a look at how it changes over time. In [1], to determine when a tweet was an advertising tweet, they simply checked whether the tweet contains a URL. By checking the content of tweets, we found that many tweets which contained URL were not ads. It is not uncommon for twitter users to include URL in their tweets. Here, a classifier was trained to filter out ads. Since the accuracy of classification is important to later data analysis, a discussion of classification has been included. The process of tweets classification as follows:

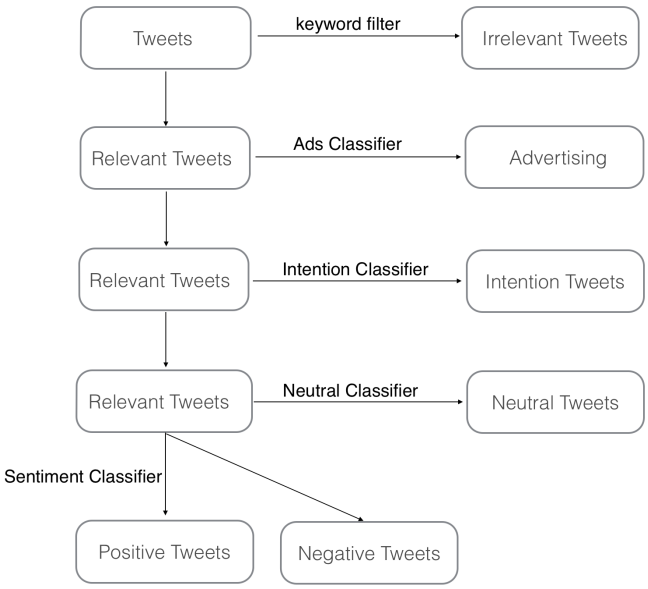


Fig 2: Tweet Classifiers

4.1 Tweets Filtering

In this project, tweets of three movies are collected. The first two movies are “Focus” and “the Lazarus effect” and collected tweets are from Feb 27 to Mar 10. The last one movies are “Cinderella” and collected tweets are from Mar 14 to April 4. At last, nearly 2.4 million tweets are collected. For each record, we only keep id, text, time, time zone and location. And tweets are stored in json format.

After the tweets were collected, we use a filtering program to filter the irrelevant tweets. In this project, we use two-step approaches. First, we used a dictionary which contains some words are relevant to movie, for example, movie, cinema, film and theatre. Also this dictionary contains some words like the Director’s name or the actors’ names. And if the tweets has both movie name and one or more words in this dictionary we think this tweet is relevant and others is not relevant as we think. Second, we build a customized dictionary for some phrase and if the tweet has one or more phrase in this dictionary, we consider this tweet is irrelevant.

Take “Focus” as an example. At first, we select tweets which contain “focus”, however, in these tweets there exist many noises. So we select again from these tweets which also contain one or more words in the dictionary we build before. Applying this approach, we receive the new set of tweets. When we go through the tweets’ content which come from new set, we find that if the tweet is irrelevant, it always appears “on” after “focus”, so we add “focus on” in the second dictionary, and then if the tweet contain the “focus on”, we consider it is irrelevant.

For ad filtering, first we need to collect a set of data which all tweets in it are advertising. Then we find a user named “movie”. This user’s tweets are almost advertising and all relevant to the movie. So we select first 1000 tweets from this user’s tweets as the dataset just contain advertising tweets. And we also manually get 500 tweets from the irrelevant tweets which are not advertising. We contain these 1500 tweets as the

training set, and use the same method we get 500 tweets as testing set which contains 350 advertising tweets and 150 no-advertising tweets. Then we train multinomial naive Bayes classifier to predict which tweets are advertising tweets.

4.2 Tweets Classification

Tweets need to be classified into four categories (intention, positive, negative and neutral). In order to apply machine learning on text documents, feature vectors are first extracted from the content of tweets. We build a bag of words from collected tweets. In bag-of-words model, a text is represented as a list of words, ignoring grammar and word order. And then with feature vectors, we can train a classifier to predict the categories of tweets. Three machine learning algorithms are evaluated: Multinomial Naive Bayes, Support Vector Machine, K-nearest neighbors. The ideas behind these three algorithms are quite different and they show different performance.

K-Nearest Neighbors algorithm is a non-parametric method used for classification and regression which means it does not make any assumptions on the underlying data distribution. This important property make KNN algorithm become more popular since most data in real world does not obey all typical assumptions. The training data would be put in an n-dimensional feature space. Then label them into different classes. The training phase of the algorithm consists only the feature vectors and the class labels of the training data. The number k is provided by users, this number decides how many neighbors influence in the classification. KNN need to find the closest k neighbors for one unknown vectors according to the distance between them. The smaller distance, the “closer” the two vectors. After finding k neighbors, vectors from the same class are grouped into sets.

Support vector machines (SVMs) are supervised learning models which always use on classification and regression analysis. The purpose of SVMs is to put all the points to the correct category by adding dimensions. Users give an SVM a set of training data, each labeled to one of two categories. The SVM training algorithms would build a model that assigns new data set to one category or the other. If some elements cannot be classified in 2D, SVM would predict the elements to higher dimensions (for example: 3D) and try to classify all the elements into the correct groups.

Multinomial Naïve Bayes model is widely used in document classification. Compared with other models, rather than taking care of the order of the words, MNB only consider the word occurrences from the document, more specifically, MNB is doing word counting. This approach has another name called “unigram language model” in the statistical language modelling for speech recognition [7]. Capturing frequency information of some token may be more helpful in some situation. For example if users want to analyze a movie, they want to know how many people would come to see this movie. They can use MNB to capture the occurrence of some intension words such as “come” , ”see” , ”want”, etc. Through the occurrences of these words to calculate the probability and finally to predict the number of viewers.

4.3 Panel Data Analysis

Panel data contains observations in different time periods for the same object like movie, individuals and it is also known as cross-section time-series data [8]. “The fundamental advantage of a panel data set over a cross section is that it will allow the researcher great flexibility in modelling differences in behavior across individuals”[8]. Panel data has different types: short panel & long panel, balanced panel & unbalanced panel. “A short panel has many entities but few time periods, while a long panel has many time periods but few entities” [9]. For a balanced panel, all entries have observations in every time period. In this project, we have a long, balanced panel data as follows:

ids	day	intention	pos	neg	netural	total	revenue
2	1	0.1479	0.6402	0.0183	0.1936	14019	6451899
2	2	0.1639	0.7108	0.0215	0.1038	7736	7634350
2	3	0.1514	0.6828	0.0217	0.1441	7746	4598888
2	4	0.1557	0.7268	0.0295	0.0879	4470	1518217
2	5	0.1006	0.7649	0.0441	0.0904	5465	1860439
2	6	0.1347	0.7440	0.0346	0.0867	2976	1286208
2	7	0.1218	0.7494	0.0347	0.0941	3975	1203495
2	8	0.1226	0.7108	0.0471	0.1195	3631	2901180
2	9	0.0904	0.8070	0.0228	0.0798	5477	4439711
2	10	0.0730	0.8118	0.0357	0.0795	6154	2666385
2	11	0.0961	0.7608	0.0470	0.0961	3767	863161
2	12	0.1024	0.7674	0.0425	0.0877	3319	1144105
1	1	0.2583	0.5514	0.0087	0.1817	10737	3786762
1	2	0.3331	0.4375	0.0191	0.2104	8386	4365431
1	3	0.2911	0.4266	0.0253	0.2570	5930	2051244
1	4	0.4283	0.4148	0.0227	0.1341	1849	567968

Fig 3: Panel Data(part)

‘ids’ represents different movies. Instead of using absolute number, we use related ratios which are the ratios of tweets with corresponding sentiment in one movie’s tweets in one day.

In this project, a fixed effect model is used to analyze panel data. “It examines if intercepts vary across over group or time period” [8]. It has functional form as follows [8]:

$$y_{it} = x'_{it}\beta + \alpha_i + \varepsilon_{it}$$

y_{it} is a variable observed for item i at time period t. x_{it} has K regressors and a_i represents all observable effects. The last item is the error term. And the least squares dummy variable model is used for estimating this fixed effect model.

5. Experiment and Design

5.1 Classifier Evaluation and Comparison

To classify intention tweets, an intention tweet dataset is built and around 3000 tweets are manually labeled. SentiStrength[10] estimates the strength of sentiment in short text and contains 4242 manually labeled tweets. In this dataset, if the strength of positive and negative sentiment are equal, we consider this tweet is neutral. And this dataset is used to train a neutral classifier. To do sentiment analysis, Sentiment140[11] is used 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.

Three different classifiers are trained and evaluated on Sentiment140[11]. From the chart below, we can find that multinomial naive Bayes has the best performance and in practice it needs much less time than other classifiers. Therefore, multinomial NB is used in following project.

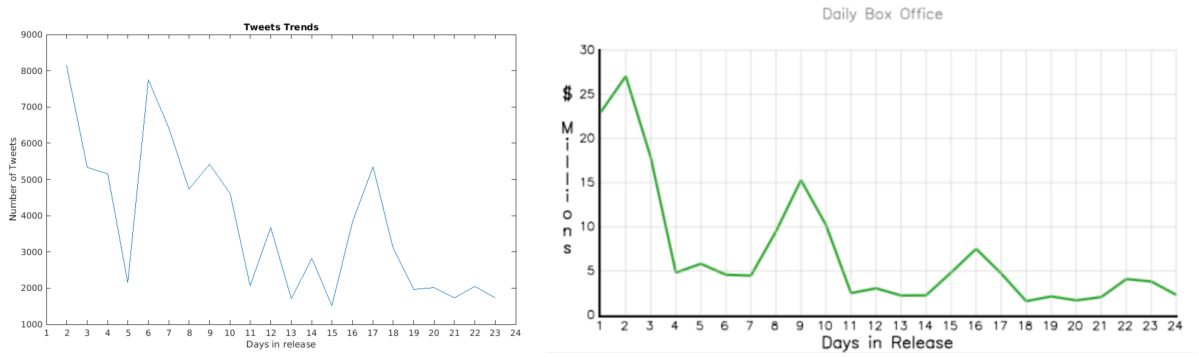
Classifier	Accuracy	Recall	f1-score
SVM	0.77	0.72	0.71
Multinomial NB	0.80	0.80	0.80
KNN	0.76	0.76	0.76

In this project, we have four dataset for text classification: one for classifying Ads, one for classifying intention tweets, one for classifying neutral tweets, and the last one for classify positive and negative tweets. For first three datasets, 90% tweets are used as training data and the rest are used as testing data. For the last dataset, classifier is tested on 500 manually labeled tweets.

Classifier		Accuracy
Ads Classifier		0.983
Intention Classifier		0.982
Neutral Classifier		0.78
Sensitive Classifier	Positive	0.82
	Negative	0.77

5.2 Tweets Analysis
5.2.1 Analysis of Tweets Statistics

In this project, three movies' tweets are collected and analyzed. The left graph shows the tweets trend for movie "Cinderella" in 23 days. The right graph shows that "Cinderella"'s movie sales and it comes from BoxOfficeMojo.com. Some patterns are observed in these two graphs. In most times, the number of daily tweets and daily movie sales have very similar changing trend. In other words, individuals tend to discuss movies on Twitter before or after they watched movies. When Friday and weekend come, the tweets number goes up and when work days back, it fall down. It is easy to understand that individuals tend to relax themselves in the weekend, which means movie sales and tweets tend are correlated.



(BoxOfficeMojo.com)

Fig 4 shows that the percentage of intention, positive, neutral or negative tweets in all film-related tweets. Intention tweets comprised 25.92% of all tweets, negative tweets 6.15%, neutral tweets 13.69% and positive tweets 54.25%.

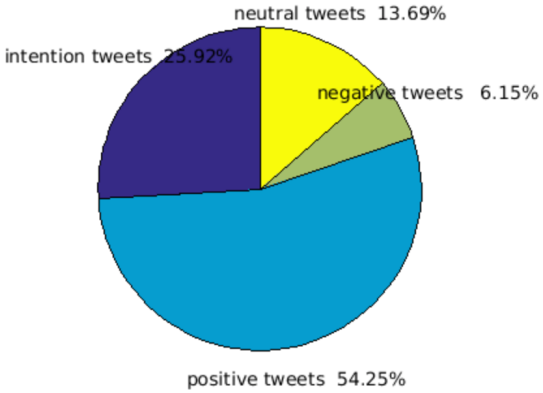


Fig 4: Percentage of Tweets

The chart below shows the average daily tweets of each movie. 'Focus' has more positive tweets than others and 'Cinderella' has more intention tweets.

Movie (Average Daily Tweets)	Intention Tweet	Positive Tweets	Negative Tweets	Neutral Tweets
Focus	728	4150	170	680
The lazarus effect	968	1531	61	585
Cinderella	1027	2150	244	543

5.2.2 Impact of Tweets on Movie

Tweets are classified into four categories and their influences on movie sales are studied separately. A fixed effect model is used to analyze panel data. From the chart below, we found that intention tweets have more positive impact on movie sales than positive tweets. The total amount of tweets also have influence on movie sales. In other words, more tweets correspond to more movie sales. Since p-value is less than 0.05, the model is right.

	Intention tweets		Positive tweets		Total tweets	
	Coefficient	Std. error	Coefficient	Std. error	Coefficient	Std. error
Daily $Tweets_{1\text{-}12}$	1867.65	347.84	839.64	131.25	466.864	75.594
R^2	0.57853		0.66089		0.64492	
p-value	2.5214E-05		2.4263E-06		3.9755E-06	

The influence of negative tweets on movie sales is also studied. Since negative tweets compose a small part of the entire tweets, we studied the negative tweets ratio to avoid the volume effect. From the chart below, we found that negative tweets have negative impact on movie sales.

	Negative tweets	
	Coefficient	Std. error
Daily $Tweets_{1\text{-}12}$	-125670112	33699777
R^2	0.39839	
p-value	0.0012393	

5.3 Conclusion

In this project, we investigated the relationship between twitter moods and movie sales. And we got following conclusions. Individuals tend to watch movies in weekend and would like to discuss movies before or after they watch movie. Positive tweets are always related to higher movie sales and negative tweets are always related to lower movie sales. Tweets, whose authors express intention to watch one movie, have more influence on movie sales than positive tweets. These conclusions are consistency to the results in previous work.

7. Reference

[1] 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.

[2] Cha, Meeyoung, et al. "Measuring User Influence in Twitter: The Million Follower Fallacy." *ICWSM* 10.10-17 (2010): 30.

[3] Bollen, Johan, Huina Mao, and Xiaojun Zeng. "Twitter mood predicts the stock market." *Journal of Computational Science* 2.1 (2011): 1-8.

[4] Tumasjan, Andranik, et al. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment." *ICWSM* 10 (2010): 178-185.

[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).

[6] 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.

[7] Andrew McCallum, Kamal Nigam (1998). A Comparison of Event Models for Naive Bayes Text Classification. *AAAI/ICML-98 Workshop on Learning for Text Categorization*, pp. 41-48. Technical Report WS-98-05. AAAI Press. 1998.

[8] Park, Hun Myoung. "Practical guides to panel data modeling: a step-by-step analysis using stata." Public Management and Policy Analysis Program, Graduate School of International Relations, International University of Japan (2011).

[9] Cameron, A. Colin, and Pravin K. Trivedi. 2009. Micro-econometrics Using Stata. TX: Stata Press.

[10] M. Thelwall, K. Buckley, G. Paltoglou, sentistrength 6humanCodedDataSets [Data set], 2014

[11] A. Go, R. Bhayani, L. Huang, sentiment140 corpus[Data set], 2009