

Using Tweets to Predict the Stock Market

Zhiang Hu, Jian Jiao, Jialu Zhu

1. Abstract

In this project we would like to find the relationship between tweets of one important Twitter user and the corresponding one stock price behavior. The tweets of Elon Musk, who is the CEO of Tesla, and the change of Tesla stock price are used as data in our project. We have tried different sets of features using SVM model.

2. Background

It is found that some “important” twitter users’ words can influence the stock prices of certain stocks. The stock price of Tesla – a famous electric automobile company – for example, recently seen a huge rise after Elon Musk, the CEO of Tesla, updated his twitter about the self-driving motors. Besides, the Dow Jones and S&P 500 indexes dropped by about one percent after the Twitter account of Associated Press falsely posted the message about an explosion in the White House.

3. Prior Work

A number of prior works are there to predict stock price using Twitter or other social network. In 2010, Bollen et al. predicted the direction of DJIA movement using Twitter data with an accuracy of 87.6% using the Google Profile of Mood States (GPOMS) and a self-organizing Fuzzy Neural Network [2010]. They found that the “Calm” mood profile yields the best result for stock market prediction [1].

4. Methods

4.1 Data Process

4.1.1 Tweet Data

One challenge of this project is to get the data from Twitter, since the large data set used in the prior work are no longer available according to Twitter’s privacy policy [2]. Therefore, we use the Twitter developer tool tweepy to obtain the content of tweets, number of “favorites”, and number of “retweets” of Twitter accounts. At most 3200 tweets information could be collected for a certain user. In this project, information of 816 tweets of Elon Musk from Jun. 4, 2010 to

Nov. 27, 2013, was collected via Tweepy.

4.1.2 Finance Data

We use the Yahoo! Finance as our data source. The finance data consists of opening price and closing price on daily basis, ranging from Dec. 1st, 2011 to Nov. 27, 2013, including 229 trading days in total. We then convert the data to the input for our model. Two kinds of labeling vector are constructed according to the finance data set: 1) we use -1 or 1 to denote price goes up or down in one trading day; 2) the feature whether there is a price jump in a specific day is used to construct the label in one of our sets of features.

4.1.3 Match of Twitter and Finance Data

We match the stock with tweets posted in the same trading day. We discard the tweets posted before stock market open time and after stock market close time. And for remaining tweets posted in the same day, we aggregate the content of tweets and average the number of “favorites” and the number of “retweets”.

4.2 Models

An SVM model with Gaussian kernel is used in this project. Three groups of features are tried and evaluated in this project in order to investigate the relationship between the tweets of an important twitter user, Elon Musk, and the direction of stock price movement of Tesla. All of these sets of features and labels are trained using SVM.

4.3 Feature selection

4.3.1 Frequent words and Jump of the stock price

The label vector of this set is to examine if the stock price changes more than $r=3\%$. The jump is labeled 1, otherwise it is labeled -1.

For the feature vector X , we observe the tweets, and use the following as features: the number of “favorites”, the number of “retweets”, and a vector constructed by the words of the tweets, which is established in the following way:

Firstly, construct a dictionary according to the content of all the tweets. Secondly, set an upper threshold of frequency F_h and a lower threshold F_l of frequency to consider only the medium frequency words. Thirdly, we compute the “Indication Factor” $\frac{P(x=i|y=1)}{P(x=i|y=-1)}$ and select k words with largest “Indication

Factors” according to the labeling vector Y. Then, the vector v is established as v_i to be the number of ith largest word appears in the tweet.

The number of “favorites” and number of “retweets” of one day are averaged and the word vectors are aggregated.

62 valid days of data points of the tweets and stock prices out of 161 stock features and 148 tweets. However, even for the training set, the error rate is quite high. Below is a figure showing the error rate under different choices of “jump”.

	Open Price	Highest Price	Lowest Price	Close Price	Volume	Adjacent Close
Error Rate (r = 3%)	25.81%	35.48%	33.87%	46.77%	6.45%	46.77%

Table 1 Error rate of different choices to construct the label vector

In table 1, for the “best case”, using volume generates 6.45% error. However, all the predictions we get are “jump”, meaning that information is lost in this hypothesis. The behavior of this group is bad even for the “jump” of the stock market. Therefore, new groups of features and labels need to be selected to examine the relationship between tweets and stock prices.

4.3.2 Sentiment scores and stock price movement

In this group of features, we use the sentiment scores of one day’s tweets to be the feature X, and the direction of stock price movement to be the label Y.

We use the following steps to preprocess the tweets for the sentiment analysis. First, the URL-like contents of tweets that contain “http” and “www” are removed. Second, the non-alpha elements such as commas in the content are deleted. Third, all the letters are turned into lower cases.

Then, we refer to prior work of other’s: we use sentiment label instead of just using the tweets itself. We use Nature Language Toolkit (NLTK) to label every tweet: positive, negative or neutral to get the sentiment scores. Table 2 is the confusion matrix of the prediction this group of features versus the real test data.

Reality\Prediction (Percentage of test sample)	Price Rises	Price Falls
Price Rises	51.52%	0
Price Falls	45.42%	3.06%

Table 2 Confusion matrix

We use all samples, 229 trading days in total, to train our model and then investigate the training error. As one could observe in Table 2, although the accuracy of our prediction is around 54%, the model tends to simply predict every test data as “price rise”, which is not acceptable. One explanation of such result is that the data in these features are inseparable, which makes the SVM model result in high bias and thus under fits the data, simply predicting almost everything as “rise”. Hence, more or some other features are needed so that the data set is more separable.

4.3.3 Time processing

In this tempt, we decide to adopt more features. First, add the number of favorites and number of retweets as our features, since more favorites and retweets indicates the tweet has more influences on the public, which will probably affect the stock price movement.

Then, we consider time series effect in our model – the most recent days’ data, because tweets effect may not influence the stock price movement immediately. We use a parameter N to denote how many days will be considered in generating a sample. And in the test we will try N from 1 to 5.

In validation part, we set a pole denoting the separating day. We use the data before separating day (inclusive) as training data, and use the data after the separating day as testing data. In the test, we found $N = 3$ is the best value. Figure 1 and Table 3 shows the accuracy and confusion matrix of this set of features and label vector.

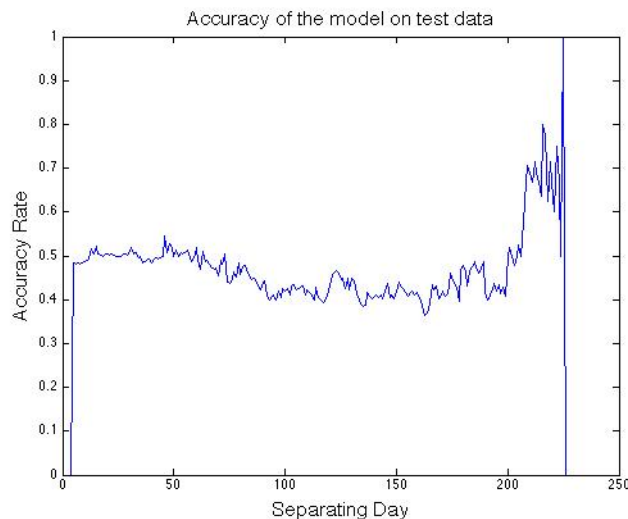


Figure 1 Accuracy of the model on test data

Reality\Prediction (Percentage of test sample)	Price Rises	Price Falls
Price Rises	15.79%	36.84%
Price Falls	5.26%	42.11%

Table 3 Confusion matrix of time processing

As presented in Figure 1 and Table 3, both the accuracy and the confusion matrix of this set of features and labeling are reasonable. Around 60% accuracy can be reached if we leave 10% data to be testing set. Besides, the confusion matrix indicates that this set of features is not resulting in predicting all rise or all fall.

5. Conclusion and future work

In this project we have tried three pairs of features and labels to investigate the relationship of an important twitter account's tweets and the stock price behavior: 1) the most frequent words and the "jump" of price, 2) the sentiment scores and the direction of movement, and 3) most recent days' aggregation of sentiment, favorites, retweets, and previous stock price movement directions. We found that only the third pair reports both reasonable accuracy and reasonable confusion matrix.

In the future, the relationship between the tweets of some important twitter accounts and the whole stock market would be investigated using the generalization of this project. Besides, features such as the tweets of the followers of these accounts could also be considered to test the relationship.

6. References

- [1] Bollen, J., Mao, H. and Zeng, X.-J. 2010. Twitter mood predicts the stock market. *Journal of Computational Science* 2(1):1–8.
- [2] Twitter, "Twitter Privacy Policy", [online] October 2013, <https://twitter.com/privacy> (Accessed 8 November 2013).