CFRM 521 Final Project

Yohan Min & Peiyuan Song

June 4, 2018

Predicting Stock Price Movement Using Social Media Analysis

Abstract

The purpose of this final project is to understand the research and replicate methods used in Derik Tsui's paper. We tried the 3 methods, Naive bayes, Support vector regression and K-nearest neighbors regression to compare with the original analysis. It turns out the results are slightly different due to the fact that the data for these analyses may be different we assume.

Background

Today's social media platforms is one of the most fast and efficient way that transport information among people as well as the financial market. This brings us an idea of study the relationship between people's comments on social media platforms and same period financial market performance. Some platforms like StockTwits provide a good and detailed resource of comments database, which we can build our model directly based on it.

Because of more and more financial institutions now start to adapt automated statistical techniques in their daily trading practices, the sample paper believes that further development of large-scale social media analysis will have the potential to introduce an additional source of investment alpha.

Same as our sample paper, our underlying assumption is that a correlation between aggregated sentiment indicator and the market price reaction. Thus, our StockTwits data which represent market sentiment can provide a robust and meaningful information of real financial market situation.

Method

Data

We first download data from StockTwits, the data is only available in JSON format, and contains more than 566,000 comments data, covering 1592 stocks from the beginning of

2013 all the way through the end of 2016. Each data point is comprised of message body, timestamp, sentiment, and ticker symbols.

We will first need to convert our raw data into txt or csv format for the ease of process with R. By doing so, we will use jsonlite package and convert raw data into matrix format. During this process, we will remove all message body, and only leave useful information in our processed sentiment dataset.

Price data for DJI Average can be retrieved using quantmod package, where we will get four years of pricing data from 2013 - 2016. Calculating forward 3-day return using exactly same method described by the sample paper. Reason of using 3-day return is to smooth out short-term volatility and market noise.

Instead of using bag of words described in the sample paper, we will use number of bullish and bearish message in a single day. For example, if in a single day, bullish comments toward DJI's constituent stocks is more than that of bearish comments, we will identify that day as a day with bullish sentiment, and vice versa. That is, sentiment parameter = 1, when (# of bullish message) > (# of bearish message).

Table 1: Cleaned data table (first 15 rows)

DJI.Close	MMM	AXP	AAPL	BA	CAT	CVX	CSCO	KO	DIS	XOM	GE	GS	HD	IBM	INTC	JNJ	JPM	MCD	MRK	MSFT	NKE	PFE	PG	TRV	UTX	UNH	VZ	V	WMT
0	0	0	1	0	0	0	1	1	0	1	0	1	0	1	1	0	1	0	0	1	0	0	0	0	0	0	0	1	1
0	0	0	1	1	1	0	0	1	0	0	1	1	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	1	0
0	0	0	1	0	1	0	1	1	1	0	1	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0
1	0	1	1	0	1	0	0	0	0	0	0	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0
1	0	1	1	1	0	0	0	0	0	0	1	1	0	1	1	0	0	1	0	1	0	1	0	0	0	0	0	1	0
1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	1	1
1	0	0	1	0	0	1	0	0	0	1	0	1	1	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0
1	0	0	1	0	0	1	0	0	0	0	1	1	1	1	1	0	0	0	0	1	1	0	0	1	0	0	0	0	1
1	0	0	1	0	0	1	1	0	0	1	1	1	0	0	1	1	1	0	0	1	0	1	0	0	0	0	0	0	0
1	0	0	1	0	1	0	1	1	0	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	1	1
1	0	1	1	0	0	0	0	0	0	0	1	1	0	0	1	0	1	0	0	1	1	0	0	0	0	0	0	0	1
1	0	0	1	0	0	0	0	0	1	0	1	1	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1
1	0	0	1	0	1	0	0	0	0	0	1	1	0	1	1	1	0	0	0	1	0	0	0	0	0	0	1	1	1
1	0	0	1	0	0	0	0	0	1	0	1	1	0	1	1	0	1	1	0	1	0	0	0	0	0	1	0	0	0

Training

Next, we will try methods described in the sample paper and try to replicate similar results. Similar to the method as described in sample paper, we will avoid look ahead bias by set 1st 75% of historical data as training data, and the remaining 25% of data as the test set (data from Jan 2016 - Dec 2016). Trainings are all performed in dinfferent methods: NB, SVR and KNN. The parameters after each train are applied to test data to compare with the real value.

Results

Naive Bayes Analysis

Table 2: NB train results in order of MMM,AXP,AAPL,BA,CAT,CVX,CSCO,KO, DIS,XOM,GE,GS,HD,IBM,INTC,JNJ,JPM,MCD,MRK,MSFT,NKE,PFE,PG,TRV,UTX,UNH,VZ,V, WMT

	0	1
mean	0.1550633	0.1750000
sd	0.3625391	0.3803996

	0	1
mean	0.4303797	0.4045455
sd	0.4959146	0.4913625

	0	1
mean	0.9588608	0.9386364
sd	0.1989272	0.2402693

	0	1
mean	0.6455696	0.6636364
sd	0.4790990	0.4730028

	0	1
mean	0.3291139	0.375000
sd	0.4706367	0.484674

	0	1
mean	0.4430380	0.4454545
sd	0.4975326	0.4975816

	0	1
mean	0.6708861	0.6181818
sd	0.4706367	0.4863854

	0	1
mean	0.4620253	0.4863636
sd	0.4993466	0.5003830

	0	1
mean	0.7879747	0.7840909
sd	0.4093910	0.4119199

	0	1
mean	0.4810127	0.4340909
sd	0.5004318	0.4962011

	0	1
mean	0.6234177	0.5431818
sd	0.4852972	0.4986989

	0	1
mean	0.5696203	0.5818182
sd	0.4959146	0.4938218

	0	1
mean	0.4683544	0.4340909
sd	0.4997890	0.4962011

	0	1
mean	0.5063291	0.5545455
sd	0.5007529	0.4975816

	0	1
mean	0.6424051	0.6431818
sd	0.4800522	0.4796058

	0	1
mean	0.4462025	0.4022727
sd	0.4978858	0.4909146

	0	1
mean	0.4493671	0.375000
sd	0.4982186	0.484674

	0	1
mean	0.3607595	0.2772727
sd	0.4809825	0.4481618

	0	1
mean	0.3449367	0.3045455
sd	0.4761016	0.4607385

	0	1
mean	0.8639241	0.8250000
sd	0.3434130	0.3803996

	0	1
mean	0.6075949	0.5590909
sd	0.4890606	0.4970611

	0	1
mean	0.4778481	0.4840909
sd	0.5003013	0.5003157

	0	1
mean	0.2943038	0.2568182
sd	0.4564520	0.4373755

	0	1
mean	0.0949367	0.0727273
sd	0.2935924	0.2599839

	0	1
mean	0.1867089	0.2068182
sd	0.3902957	0.4054850

	0	1
mean	0.306962	0.2363636
sd	0.461965	0.4253317

	0	1
mean	0.4525316	0.4909091
sd	0.4985311	0.5004864

	0	1
mean	0.6740506	0.6636364
sd	0.4694719	0.4730028

	0	1
mean	0.4683544	0.4795455
sd	0.4997890	0.5001501

Support Vector Regression Analysis

```
##
## Call:
## svm(formula = DJI.Close ~ ., data = dat_train)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  1
##       gamma:  0.03448276
##
## Number of Support Vectors:  706
##
##  ( 316 390 )
##
##
## Number of Classes:  2
##
## Levels:
##  0 1
```

K-Nearest Neighbors Regression Analysis

Table 3: Test accuracy, compared KNN with 1,5, and 20

	Test Accuracy
KNN1	0.5020
KNN5	0.4538
KNN20	0.5141
Average	0.4900

Summary & Discussion

Among the 3 methods, SVR and KNN perform better than NB with the test accuracy of 51.81% an 50.2% respectively. NB has the accuracy of 49.4%. These results are a bit different from the original anaysis that shows the results of accuracy for NB, SVR and KNN as 50.9%, 56.82% and 54.48%.

Table 4: Test accuracy, compared

	Test Accuracy
Naive.Bayes	0.4940
SVR	0.5181
KNN	0.5020
Average	0.5047

Also with respect to the individual stock predicting the market of stock price movement, our result shows that CAT has the highest test accuracy as 63.5% while AAPL has the lowest test accuracy of 30.5%. The overall discrepency between the original analysis and our results may be due to the difference of data used for analyzing the stock price movement. Since we can't figure out how the data is different and how the data the orginal analysis is based on was processed and cleaned, we are just guessing. But in general, we also found that NB performed less than the other two methods (i.e. SVR and KNN).

Sentiment in social media could be the good pedictor to forecast the market price movement. Due to the improvement of computational power and analysis techniques it is possible to estimate the uncertain market behavior better than before. On the other hand, there is still a limit as there are always uncertain phenomena that is hard to catch. Although Our result shows that the test accuracy is slightly betther than 50%, including other statistical methods to improve the processes that we used in this report, will enhance the test accuracy.

Another limit from the database we are using is that messages are already identified as bullish and bearish by data provider. Some of messages are clear in their sentiment toward financial market while there are some of them can be considered as quite neutual. In that case, whether to catigorized them under bullish or bearish can be a very subjective decision and will produce bias. One way to improve is to consider add another category of neutral comments, so that we can collect all neutral and ambiguous comments under this category.

Table 5: Test accuracy, compared

	Test accuracy(%)
MMM	57.8
AXP	52.6
AAPL	30.5
BA	47.4
CAT	63.5
CVX	47.8
CSCO	37.3
KO	47.8
DIS	40.6
XOM	43.8
GE	38.6
GS	45.8
HD	46.2
IBM	56.6
INTC	39.8
JNJ	44.6
JPM	43.8
MCD	33.3
MRK	41.8
MSFT	30.9
NKE	37.3
PFE	44.2
PG	43.4
TRV	54.2
UTX	58.2
UNH	41.4
VZ	51.4
V	44.6
WMT	45.0

References

1. Derek G. Tsui. Predicting Stock Price Movement Using Social Media Analysis, Stanford University, 2016