Prediction of Stock Market with Sentiment Analysis

Lin Zeng(lz447), Qin Lu(ql224), Sizhang Zhao(sz459)

October 27, 2016

1 Idea

Stock market prediction is the core research area in trading and investment. Stock price is determined by the behavior of human investors, and the investers determine stock prices by using publicly available information to predict the market future trend. Financial news can thus play a significant role in influencing the stock trend as human react to the information. Previous reasearch have suggested that there is some lag between the time news article was released and the time the market absorbed and reflected these information. So our main goal here is to classify the news as our sentiment factor and use the sentiment factor to determine the impact of the news on the stock price.

In our research, we focus on sentiment analysis with Google news in stocks division, and build up machine learning models to measure the relationship between news sentiment and stock index trend(DJIA) over time. Finally, we try to design a tradable portfolio based on our sentiment factor and test its application in investments.

2 Data Scraping

Headers of financial news usually contain the impact of a certain event to the whole market. For example, *warn, volatility, decline, plunge* are common phrases that are conceived as negative news to market, and *hike, allow, optimism, gain* are phrases that are considered as positive ones. So in order to capture as much sentiment information as possible for news each day, we use the headers of financial news as our data set instead of the whole article.

We designe a web scraper to scratch the Google news headers in stock division from January 2007 - December 2009 and January 2013- December 2015 in daily basis.

By locating the headers in source code from Google news, we use *BeautifulSoup* package in Python to scratch the headers in the search page by changing the time zone in advanced search. We store news headers scratched from Google news each day in separate *.txt* file as the raw data for sentiment analysis.

In addition, we alter the news search criterion by industry to predict the specific industry trend. For example, we search *real estate* instead of *stock* to obtain specific industry news in predicting real estate index future trend.

We also download the *Dow Jones Industrial Average* (DJIA) as our response variable data from Yahoo Finance.

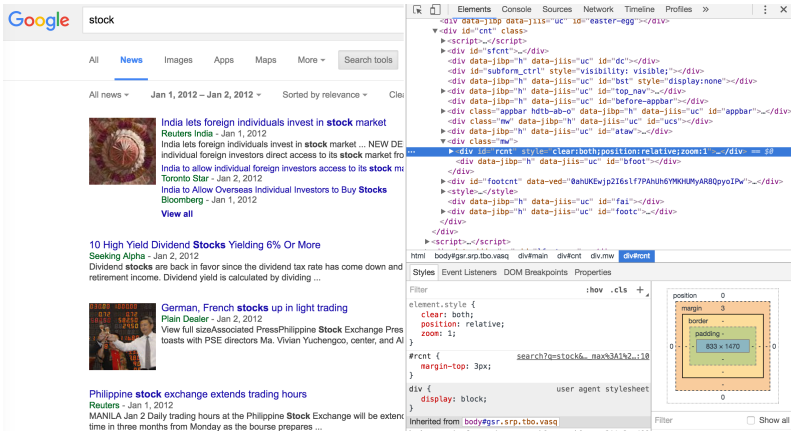

Figure 1: Source code and the search page of Google news

Figure 2: News headers stored in .txt file by our web scraper

3 Sentiment Analysis with Stock News

Our goal here is to develop a system capable of providing information about the polarity in our scratched news documents. *OpinionFinder*(OF) is a publicly available software package that processes documents and automatically identifies subjective sentences and sentiment expresssions. The subjectivity analysis provides us three polarity level: *positive*, *neutral*, *negative*.

The polarity classifier takes clues consisting of words with a prior polarity of *positive*, *negative* or *neutral* (for example, *hike*, *drop*, *indicate*, repectively) and thus uses a modified version of the classifier described in Wilson et al. (2005) to determine the contextual polarity of the clues. Heuristics were used to improve the speed of the classifier so it no longer needs the dependency parse output. When evaluated on the MPQA opinion corpus, the overall accuracy is 73.4%.

After obtaining the polarity word list, we mark positive words as 1, neutral as 0 and negative as -1, and take the average of them as our sentiment factor. The sample output of OF is demonstrated as follow.

<MPQASENT autoclass1="unknown" autoclass2="obj" mpqapolarity= "neutral">We have no record of Mark Twain's earliest letters.</MPQASENT>

<MPQASENT autoclass1="subj" autoclass2="subj" mpqapolarity= "neutral">Very likely they were soiled pencil notes, written to some school sweetheart--to "Becky Thatcher," perhaps--and tossed across at lucky moments, or otherwise, with happy or disastrous results.</MPQASENT>

<MPQASENT autoclass1="unknown" autoclass2="subj" mpqapolarity= "positive">One of those smudgy, much-folded school notes of the Tom Sawyer period would be priceless to-day, and somewhere among forgotten keepsakes it may exist, but we shall not be likely to find it.</MPQASENT>

<MPQASENT autoclass1="unknown" autoclass2="subj" mpqapolarity= "negative">No letter of his boyhood, no scrap of his earlier writing, has come to light except his penciled name, SAM CLEMENS, laboriously inscribed on the inside of a small worn purse that once held his meager, almost non-existent wealth.</MPQASENT>

Figure 3: Sample output of OpinionFinder

4 Granger Causality Analysis

We are concerned with the question whether the changes in our sentiment factor is correlated with changes in the stock market, in particular DJIA $\prod$closing values. To answer this question, we apply the econometric technique of Granger causality analysis to the daily time series produced by sentiment factor and the DJIA. Granger causality analysis rests on the assumption that if a variable X causes Y then changes in X will systematically occur before changes in Y. We will thus find that the lagged values of X will exhibit a statistically significant correlation with Y. We are not testing actual causation but whether one time series has predictive information about the other or not.

Our DJIA time series, denoted as D_t, is the adjusted close price of DIJA at time t. To test whether our sentiment factor time series predicts the changes in stock market value, we compare the variance explained by two linear models as show below. The first model (L_1) uses only the lagged value D_{t-1}, while the second model (L_2) uses the lagged value D_{t-1} and the 3 lagged values of sentiment factor $X_{t-1}, X_{t-2}, X_{t-3}$.

$$L_1 : D_t = \alpha + \sum_{i=1}^{3} \beta_i D_{t-i} + \varepsilon_t$$

$$L_2 : D_t = \alpha + \sum_{i=1}^{3} \beta_i D_{t-i} + \sum_{i=1}^{3} \gamma_i D_{t-i} + \varepsilon_t$$

The results are shown below indicating the adjusted R^2 of L_2 is larger than L_1, this means Granger causality analysis suggests a predictive relation between certain sentiment dimensions and DJIA.

L1 regression report

Dep. Variable:	y	R-squared:	0.819
Model:	OLS	Adj. R-squared:	0.805
Method:	Least Squares	F-statistic:	55.87
Date:	Thu, 03 Nov 2016	Prob (F-statistic):	8.18e-14
Time:	16:18:15	Log-Likelihood:	-23.117
No. Observations:	41	AIC:	54.23
Df Residuals:	37	BIC:	61.09
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
const	4.06e-15	0.070	5.81e-14	1.000	-0.142 0.142
x1	0.9803	0.166	5.923	0.000	0.645 1.316
x2	0.0267	0.233	0.115	0.909	-0.446 0.499
x3	-0.1248	0.170	-0.732	0.469	-0.470 0.221

Omnibus:	7.485	Durbin-Watson:	2.024
Prob(Omnibus):	0.024	Jarque-Bera (JB):	6.227
Skew:	-0.825	Prob(JB):	0.0445
Kurtosis:	3.961	Cond. No.	6.83

Figure 4: L_1 regression summary

L2 regression report

Dep. Variable:	y	R-squared:	0.871
Model:	OLS	Adj. R-squared:	0.848
Method:	Least Squares	F-statistic:	38.20
Date:	Thu, 03 Nov 2016	Prob (F-statistic):	1.02e-13
Time:	16:18:15	Log-Likelihood:	-16.223
No. Observations:	41	AIC:	46.45
Df Residuals:	34	BIC:	58.44
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
const	1.915e-15	0.062	3.11e-14	1.000	-0.125 0.125
x1	0.8412	0.153	5.499	0.000	0.530 1.152
x2	0.1656	0.210	0.791	0.435	-0.260 0.591
x3	-0.1983	0.152	-1.307	0.200	-0.507 0.110
x4	0.2311	0.072	3.195	0.003	0.084 0.378
x5	-0.0541	0.071	-0.757	0.455	-0.199 0.091
x6	0.2135	0.075	2.844	0.007	0.061 0.366

Omnibus:	1.493	Durbin-Watson:	2.027
Prob(Omnibus):	0.474	Jarque-Bera (JB):	1.300
Skew:	-0.422	Prob(JB):	0.522
Kurtosis:	2.778	Cond. No.	7.23

Figure 5: L_2 regression summary

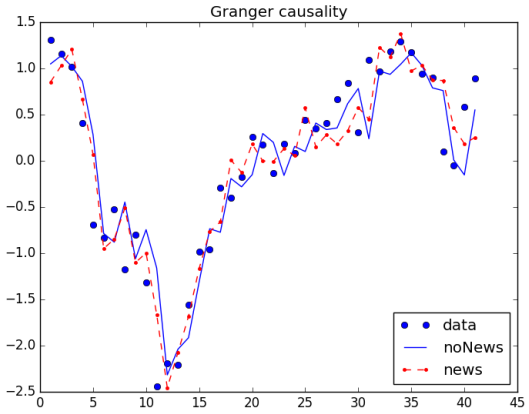

Figure 6: L_1, L_2 prediction of DJIA and true DJIA (normalized)

5 Next Step

Our Granger causality analysis suggests a predictive relation between certain sentiment dimensions and DJIA. However, Granger causality analysis is based on linear regression whereas the relation between the market sentiment and stock market value is almost certainly non-linear. To better addree these non-linear effects and assess the contribution that sentiment factor can make in predictive models of DJIA values, we will compare the performance of a *Self-organizing Fuzzy Neural Network* (SOFNN) model that predicts DJIA values on the basis of two sets of inputs:

- The past 3 days of DJIA values
- the same combined with the sentiment factors

Statistically significant performance differences will allow us to either confirm or reject the null hypothesis that market sentiment measurement do not improve predictive models of DJIA values.

We use a SOFNN as our prediction model since they have previously been used to decode nonlinear time series data which describe the characteristics of the stock market and predict its values. Our SOFNN in particular is a five- layer hybrid neural network with the ability to self-organize its own neurons in the learning process.

Having predicted the DJIA closing values one day in advance, we can use these predicted values to make sell/buy decisions. We develop a naive greedy strategy based on a simple assumption that we can hold at most one stock at any given time. Following are steps of our strategy(to be tested):

- Pre-computation

  We maintain a running average and standard deviation of actual adjusted stock values of previous k days
- Sell Decision

  If the predicted stock value for the next day is n standard deviation less than the mean, we sell the stock else we hold.
- Buy Decision

  If the predicted stock value is m standard deviation more than the actual adjusted value at buy time, we buy the stock else we hold.