Correlating S&P 500 Stocks with Twitter Data

Yuexin Mao
University of Connecticut
yuexin.mao@uconn.edu

Bing Wang
University of Connecticut
bing@engr.uconn.edu

Wei Wei
University of Connecticut
weiwei1@gmail.com

Benyuan Liu
University of Massachusetts,
Lowell
bliu@cs.uml.edu

ABSTRACT

Twitter is a widely used online social media. One important characteristic of Twitter is its real-time nature. In this paper, we investigate whether the daily number of tweets that mention Standard & Poor 500 (S&P 500) stocks is correlated with S&P 500 stock indicators (stock price and traded volume) at three different levels, from the stock market to industry sector and individual company stocks. We further apply a linear regression with exogenous input model to predict stock market indicators, using Twitter data as exogenous input. Our preliminary results demonstrate that daily number of tweets is correlated with certain stock market indicators at each level. Furthermore, it appears that Twitter is helpful to predict stock market. Specifically, at the stock market level, we find that whether S&P 500 closing price will go up or down can be predicted more accurately when including Twitter data in the model.

Categories and Subject Descriptors

H.3.3 [**Information Systems Applications**]: Miscellaneous

General Terms

Measurement, Performance

Keywords

Twitter, Stock, Correlation, Prediction

1. INTRODUCTION

Twitter is a widely used online social media. The fast growth of Twitter has drawn much attention from researchers in different disciplines. Due to the real-time nature of tweets, researchers have become interested in using Twitter to predict stock market. A recent study [3] finds that specific public mood state in Twitter is significantly correlated with the Dow Jones Industrial Average (DJIA), and thus can be used to forecast the direction of DJIA changes with high accuracy. Another study [5] finds that emotional tweet percentage is correlated with Dow Jones, NASDAQ and S&P 500.

In this paper, instead of focusing on sentiment, we investigate the correlation between the daily number of tweets that mention Standard & Poor 500 (S&P 500) stocks and several S&P 500 stock indicators (that are related to stock price and traded volume). Our investigation is at three different levels, from the stock market, to industry sector, and then to individual company stocks. Our main findings are:

- We find that at the stock market level, the daily number of tweets that mention S&P 500 stocks is significantly correlated with S&P 500 daily closing price. It is also correlated with S&P 500 daily price change and S&P 500 daily absolute price change. At the industry sector level, for eight out of the ten GICS (Global Industry Classification Standard) industry sectors, there exists significant correlation between the number of daily tweets and the daily traded volume for the sector. In particular, four sectors, Financials, Energy, Health Care and Materials, show the strongest correlations. Last, at the individual company stock level, we focus on Apple Inc.'s stock, the most tweeted stock, and find that the daily number of tweets mentioning Apple Inc.'s stock is strongly correlated with the stock's daily traded volume and absolute price change. Furthermore, it is also correlated with daily price change.

- We further apply a linear regression with exogenous input model to predict stock market indicators, where Twitter data serve as exogenous input. Specifically, at each level, we identify the highly correlated pairs, and predict stock market indicators using the model. Our preliminary results show that Twitter seems to be helpful to predict stock market. Specifically, at the stock market level, we find that whether S&P 500 closing price will go up or down can be predicted more accurately when including Twitter data in the model.

Our findings on Apple Inc.'s stock are consistent with those in [4]. However, the study in [4] only considers individual company stocks, while our study also considers the other two levels, namely stock market and sector levels.

HotSocial'12 August 12, 2012, Beijing, China.

The rest of the paper is organized as follows. Section 2 describes data collection methodology and the data sets. Section 3 presents correlation and prediction results. Last, Section 4 concludes the paper and presents future work.

2. DATA COLLECTION

2.1 Stock market data

We obtained daily stock market data from Yahoo! Finance [2] for the 500 companies in the S&P 500 index from February 16 to May 10, 2012. At the stock market level, we consider four stock indicators: S&P 500 daily closing price, S&P 500 daily traded volume, S&P 500 daily price change, and S&P 500 daily absolute price change. Daily price change is the price difference between the closing prices of the current day and previous day. It can be either positive or negative, corresponding to the rise or fall of the stock price, respectively. Daily absolute price change is the absolute value of daily price change, which can help people to find out the best option trading strategies.

At the sector level, we record the daily traded volume for each of the ten GICS sectors. GICS is an industry taxonomy developed by MSCI and S&P for use by the global financial community [1]. The GICS structure consists of ten industry sectors, including Information Technology (IT), Financials, Consumer Discretionary (CD), Consumer Staples (CS), Industrials, Energy, Health Care (HC), Materials, Telecommunications Services (TS) and Utilities. S&P 500 classifies each of the 500 companies into one of the ten industry sectors. For each sector, the daily traded volume of the sector is the sum of daily traded volume of all the companies in the sector. Table 1 presents the number of companies in each sector and average daily tweets we collected for each sector.

At the company stock level, we focus on the Apple Inc.'s stock, since this stock is most tweeted (see Section 2.2). Same as that at the stock market level, we consider four indicators for this stock: the daily closing price, daily traded volume, daily price change and daily absolute price change.

2.2 Twitter data

In Twitter community, people usually mention a company's stock using the stock symbol prefixed by a dollar sign, for example, $AAPL for the stock of Apple Inc. and $GOOG for the stock of Google Inc. We use Twitter API to search for public tweets that mention any of the S&P 500 stocks using the aforementioned convention (i.e., putting a $ before the stock symbol). The reason why we use this convention is that some stock symbols are common words (e.g., A, CAT, GAS are stock symbols), and hence using search keywords without the dollar sign will result in a large amount of spurious tweets. Fig. 1 is the log-log plot of average number of daily tweets for each of the 500 stocks in descending order. On average, there are 9434 daily tweets that mention these 500 stocks. Apple Inc.'s stock is most tweeted, hence we focus on Apple Inc. at the company stock level.

We use the daily number of tweets for S&P 500 stocks as the Twitter predictor at stock market level, use the daily number of tweets for each sector as the Twitter predictor at sector level, and use the daily number of tweets that mention Apple Inc.'s stock as the Twitter predictor at the individual company stock level.

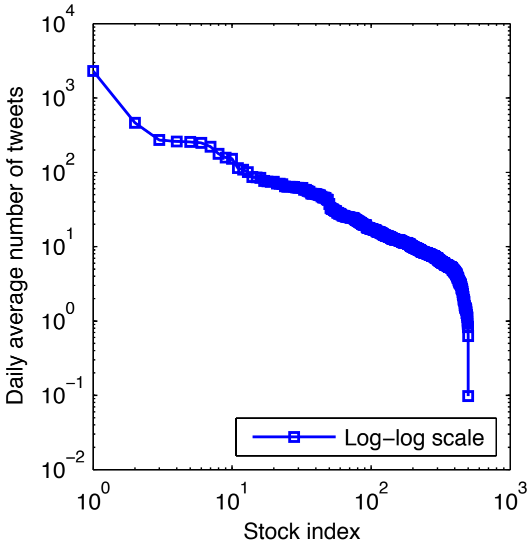

Figure 1: Log-log plot of the average number of daily tweets for each of the 500 stocks.

2.3 Data Normalization

To provide a common scale for comparison of our predictors and stock market indicators, we normalize each time series to z-scores on the basis of a local mean and standard deviation. For example, for a dataset X, the z-score of x_i in X, denoted as $Z(x_i)$, is defined as:

$$Z(x_i) = \frac{x_i - \mu(X)}{\sigma(X)} \tag{1}$$

where $\mu(X)$ and $\sigma(X)$ represent respectively the mean and standard deviation of X in a time window.

3. RESULTS

In this section, we first investigate the correlation between our Twitter predictor and the stock market indicators at each of the three aforementioned levels. After that, we select the high correlation pairs at each level, and apply a linear model to predict stock market indicators using Twitter as exogenous input.

3.1 Correlation between Tweet Volume and Stock Indicators

We use a linear regression model to test the correlation between the Twitter predictor and stock indicators at the three levels. The regression model is

$$Y_t = \alpha + \beta X_t + \varepsilon_t, \tag{2}$$

where Y_t represents a stock indicator on day t , X_t represents the related Twitter predictor on day t, α is the intercept, β is the slope, and ε_t is a random error term for day t. α and β are the regression parameters need to be determined. The whole 56 days' of data collected from February 16 to May 10, 2012 are used for correlation analysis. The correlation analysis provides three quantities: the root mean square error (RMSE), the correlation coefficient, denote as R, and the p value.

Table 1: Number of companies and average number of tweets for the ten GICS sectors.

GICS sector	Number of Companies	Average number of daily tweets
Information Technology (IT)	70	4451
Financials	81	1716
Consumer Discretionary (CD)	82	1649
Consumer Staples (CS)	41	783
Industrials	62	754
Energy	41	660
Health Care (HC)	51	540
Materials	29	406
Telecommunications Services (TS)	8	292
Utilities	35	179

**Table 2: Correlation results at stock market level: correlation between the number of daily tweets for S&P 500 and the four S&P 500 stock indicators. ***** $p < 0.01$, ***** $p < 0.1$.

S&P 500 indicators	RMSE	R	p value
Closing price	0.943	0.356	0.007***
Traded volume	1.003	0.114	0.399
Price change	0.984	0.224	0.097*
Absolute price change	0.977	0.252	0.060*

3.1.1 Stock Market Level

At the stock market level, we evaluate the correlation between the number of daily tweets and each of the four stock indicators for S&P 500 introduced in Section 2.1. The results are presented in Table 2. We see the number of daily tweets is linearly correlated with the S&P 500 closing price with a 1% level of significance, and is correlated with S&P 500 daily price change and S&P 500 daily absolute price change with a 10% level of significance.

3.1.2 Sector Level

At the sector level, we first test the correlation between the number of daily tweets and the daily traded volume for each sector. Table 3 summarizes the results, indicating a significant correlation between the number of daily tweets and the daily traded volume for eight out of the ten sectors (the other two sectors, Consumer Discretionary and Consumer Staples, have week correlation). Among these eight sectors, Financials, Energy, Health Care and Materials sectors show the most significant correlation (with a 1% level of significance).

We further analyze the correlation between the number of daily tweets of each sector and the overall S&P 500 indicators. We are interested in finding out whether sectors with a large number of daily tweets such as the Financials sector can reflect the overall S&P 500 stock information. Interestingly, the number of daily tweets of the Financials sector is correlated with both the S&P 500 daily traded volume and daily closing price with the p value of 0.041 and 0.022, respectively. Both have a 5% level of significance. Note that the Financials sector is the second most tweeted sector. On the other hand, for the most tweeted sector, Information Technology, there is no strong correlation between the number of daily tweets and the S&P 500 daily traded volume and daily closing price.

**Table 3: Correlation results at sector level: correlation between the daily traded volume and the number of daily tweets for each GICS sector. ***** $p < 0.01$, ****** $p < 0.05$, ***** $p < 0.1$.

GICS sector	RMSE	R	p value
IT	0.963	0.300	0.025**
Financials	0.796	0.614	0.000***
CD	0.988	0.206	0.127
CS	0.999	0.139	0.307
Industrials	0.964	0.298	0.026**
Energy	0.868	0.511	0.000***
HC	0.905	0.442	0.001***
Materials	0.780	0.635	0.000***
TS	0.966	0.291	0.029**
Utilities	0.979	0.241	0.073*

**Table 4: Correlation results at company stock level: correlation between the stock indicators and the number of daily tweets for Apple Inc. ***** $p < 0.01$, ****** $p < 0.05$.

Apple Inc. stock indicators	RMSE	R	p value
Closing price	1.001	0.125	0.361
Traded volume	0.760	0.658	0.000***
Price change	0.952	0.331	0.013**
Absolute price change	0.861	0.521	0.000***

3.1.3 Company Stock Level

At the individual company stock level, we investigate Apple Inc.'s stock (AAPL), which is the most tweeted stock among the S&P 500 companies. We evaluate the correlation between the number of daily tweets and each of the four Apple Inc. stock indicators introduced in Section 2.1. The results are presented in Table 4. We find that the daily traded volume as well as the daily absolute price change are strongly correlated with the number of daily tweets. Recall that at the stock market level, the number of daily tweets and daily absolute price change for S&P 500 are strongly correlated too. This seems that when stock prices change sharply during a day, no matter going up or down, more people are likely to tweet about stocks, resulting in a strong correlation between the tweet volume and the absolute price change.

3.2 Trend Prediction of Stock Indicators Using Twitter Data

After establishing that the Twitter predictors are correlated with the stock market indicators, we are interested in finding out whether and how well the stock market indicators can be predicted using Twitter data. To answer this question, we apply a linear regression with exogenous input model to our Twitter predictors and the stock market indicators. The prediction model is

$$Y_t = \alpha + \sum_{i=1}^{m} \beta_i Y_{t-i} + \sum_{i=1}^{n} \gamma_i X_{t-i} + \varepsilon_t, \tag{3}$$

where t indexes days, Y_t denotes a stock market indicator on day t, X_t denotes the number of related tweets during day t, α, β_i and γ_i are the regression parameters need to be determined, and ε_t is the random error term for day t.

In our experiments, for each level, we use the first 37 days of data as training set to build a prediction model (i.e., determine the parameters in (3)), and predict the results for the 38th day. We then use the first 38 days of data as training set to build a prediction model, and predict the results for the 39th day. This process continues for the rest of the data set (which contains data for 56 days). To build a model, the orders of the model, m and n, are unknown and must be estimated. For that purpose, we vary m from 0 to 2 and n from 0 to 3 to obtain the best combination of m and n values, and use that for prediction. Specifically, for each combination of m and n, we build the regression model based on equation (3), calculate the root mean square error (RMSE), and choose the model with the lowest RMSE. Note that, as a special case, the models obtained using $n = 0$ are models where Twitter data are not used.

At stock market level, we find that the number of daily tweets for S&P 500 is correlated with the S&P 500 daily closing price with a 1% level of significance. We now evaluate whether and how well the number of daily tweets can predict S&P 500 daily closing price. We find that for all the 19 predictions, the best prediction model has $n > 0$, indicating that including Twitter data is helpful in building more accurate prediction models. For day t, we use -1 and 1 to represent the direction of change for S&P 500 closing price from day $t-1$ to t. Specifically, 1 represents that the S&P 500 daily closing price goes up and -1 represents that the S&P 500 daily closing price goes down. Fig. 2 plots the predicted and the actual direction of change. The prediction accuracy is 68%, significantly larger than random guess. Furthermore, we observe less prediction errors from Fig. 2 for later days, which may indicate having more data in the training set is helpful to reduce errors. As future work, we will validate the results using long-term data.

At sector level, we find that the number of daily tweets of the Financials sector is strongly correlated with its corresponding daily traded volume. Hence we investigate whether and how well the number of daily tweets of the Financials sector can predict its daily traded volume. We again find including Twitter data is helpful in building more accurate prediction models. Using the same prediction method as that at the stock market level, the direction prediction accuracy is again 68%.

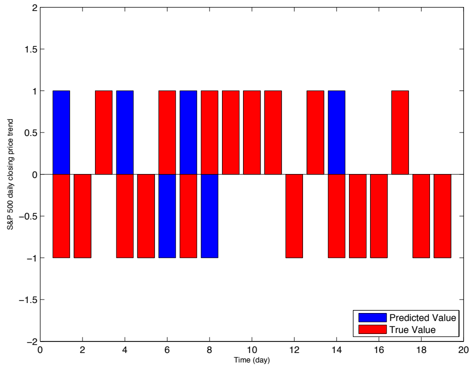

Figure 2: Predicted and actual direction of change for S&P 500 daily closing price.

At the company stock level, we observed a strong correlation between the number of daily tweets and daily traded volume for Apple Inc.. Hence we investigate whether and how well the number of daily tweets can predict the daily traded volume. However the accuracy in predicting the direction of change in daily traded volume is only 52%. Further investigation using long-term data is left as future work.

4. CONCLUSION & FUTURE WORK

In this paper, we have investigated whether the daily number of tweets that mention S&P 500 stocks is correlated with several S&P 500 stock indicators. This was done at three different levels from the stock market to industry sector then to individual stocks. We also applied a linear regression with exogenous input model to predict stock market indicators, using Twitter data as exogenous input. Our results demonstrate that daily number of tweets is correlated with the stock market indicators. Furthermore, it seems that Twitter data can be useful to predict stock market.

As future work, we are pursuing in the following four directions. First, we will consider more search words in order to collect more relevant tweets each day. Instead of only focusing on the search words of stock symbol prefixed by a dollar sign, we will add more stock related search words, and then filter out spurious tweets. Second, we are in the process of collecting long-term data, and will validate the results using long-term data. Third, we will combine both the number of tweets and sentiment results in prediction. Last, we will investigate whether and how the prediction results can help a stock trader to make trading decisions.

5. REFERENCES

[1] Wikipedia. http://en.wikipedia.org/wiki/GICS/.

[2] Yahoo! finance. http://finance.yahoo.com/.

[3] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1 – 8, 2011.

[4] E. J. Ruiz, V. Hristidis, C. Castillo, A. Gionis, and A. Jaimes. Correlating financial time series with micro-blogging activity. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 513–522, 2012.

[5] X. Zhang, H. Fuehres, and P. A. Gloor. Predicting stock market indicators through twitter, i hope it is not as bad as i fear. *Procedia - Social and Behavioral Sciences*, 26(0):55 – 62, 2011.