Enhancing Financial Decision-Making Using Social Behavior Modeling

Ruoqian Liu, Ankit Agrawal, Wei-keng Liao, Alok Choudhary
Department of Electrical Engineering and Computer Science
Northwestern University
Evanston, IL 60208
{rll943,ankitag,wkliao,choudhar}@eecs.northwestern.edu

ABSTRACT

Financial trading is a social activity that involves every participant's decision making. Meanwhile, people's online behavior collectively creates the public emotion which affects investors' reactions and hence market movements. This process can be modeled by connecting online social behavior and future trading behavior to better understand mechanisms of the stock movement so as to assist financial decision making. In this paper, we investigate the query information of financially related Wikipedia pages, and show that early signs of trading volume movements can be detected which expose financial risks. We embed this information into a classic pairs trading strategy acting on a large portfolio of stocks. Over 23% profits are seen when testing on the year of 2013 and 20% comes from the inclusion of online social data.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications - Data Mining

General Terms

Algorithm, Measurement, Performance

Keywords

Social activity modeling, Wikipedia, Algorithmic trading

1. INTRODUCTION

Understanding the stock market to predict its movement is of great interest to investors, businesses, and academia. Sometimes overlooked by researchers, investing in stock-like speculative assets is a social activity, for that the price movement solely depends on people's economic decisions and actions made to the trading market. People's decisions, even those of the most rational investors, are largely driven by effects of psychological, social, cognitive, and emotional factors, which a great portion of the uncertainty and randomness in the market is attributed to. Stock market data contain extremely detailed records of economic decisions made by traders. Besides the price data, trading volume, the number of shares transacted every day, is a more direct reflection of market interest and people's intention. The financial market and people's social behavior are closely connected and it is therefore justified to exploit factors from a social-psychological standpoint in order to make profits.

A growing body of research has explored the use of data mining methods [12] to assist stock trading. The exploration social behaviors has also been popular. Changes in Google Trend search volume [10] and Wikipedia usage patterns [9] are linked to stock market moves. Web search and query data were used by simple regression models to predict Dow Jones Industrial Average (DJIA), trading volumes and other financial indicators [7]. In [5] the author measured collective mood states from Twitter feeds and demonstrated improvement on DJIA prediction. The limitation of these works is the lack of a large enough set of stocks, or a portfolio, to minimize the risk. Although the recent work in [8] investigated Twitter volume spikes using all S&P 500 stocks, the stocks are executed rather individually.

In this work, we investigate whether the query data from the popular online encyclopedia *Wikipedia* [3] can be linked to profitable decision making in the stock market. Specifically, we ask if changes in the viewing volume of pages of companies relate to the trading volume of corresponding stocks on the next day, and what insight do both volumes provide to the price movement. The social behavior influence is incorporated into a profound trading strategy, acting on a collective portfolio of stocks to minimize the risk. This paper makes the following major contributions.

- We investigate the relationship of the *Wikipedia* query volume and the trading volume, a critical information from the stock market data, both regarding each individual stock (and its company).
- On the basis of a classic trading strategy, pairs trading, we formulate an enhanced trading strategy incorporating social activity information, and demonstrate the improvement on profitability.

The rest of the paper is organized as follows. Section 2 discusses the concept of pairs trading. Section 3 explains the financial and social data we use. Section 4 describes in detail our overall strategy and Section 5 demonstrates the results in real-world testing. Section 6 concludes the work.

2. BACKGROUND

The 8th SNA-KDD Workshop '14 (SNA-KDD '14) August, 2014, New York, NY USA

http://dx.doi.org/10.1145/2659480.2659505 .

Pairs trading [13] is a statistical arbitrage strategy, where a pair of stocks P and Q from the same industry or having similar characteristics are expected to follow similar patterns. When stock P outperforms stock Q, Q is bought long while P is sold short. The risk from the overall market is avoided as we bet on relative movements. The profitability is therefore independent from the market condition. We denote P_t and Q_t as the time series price signals of the two stocks. By taking a logarithmic form and a derivative on the returns of them over some period of time starting from t_0, we can model the system as

$$\frac{dP_t}{P_t} = \alpha dt + \beta \frac{dQ_t}{Q_t} + dX_t \quad (1)$$

where X_t is a stationary residual that oscillates around 0. In many cases of interest, the drift α is trivial. The fact of X_t being stationary indicates that the portfolio contains a linear combination of P and Q oscillating near some statistical equilibrium. Another way to understand it is that the returns of P and Q (times β) are each other's fair price and should eventually converge to the same value.

A generalized extension to pairs trading is to evolve from a pair to a portfolio to further reduce the risk. By decomposing stock returns into systematic and residual components, the equation will look like

$$\frac{dP_t}{P_t} = \alpha dt + \sum_{j=1}^{m} \beta_j F_t^{(j)} + dX_t \quad (2)$$

where the terms $F_t^{(j)}, j = 1, \ldots, m$ represent returns of some systematic factors that together determine the market's major trend. Our strategy uses Principle Component Analysis (PCA) to find the systematic factors $F_t^{(j)}$, and the resulting residuals are treated as trading signal indicators.

3. DATA COLLECTION

3.1 Stock Market Data

We obtained daily stock market data from Yahoo! Finance [4] for the top 100 large-capitalization stocks in the S&P 500 index in a time span of four years between 2010-01-04 and 2013-12-31. For each stock we consider two types of data: stock daily closing price and stock daily trading volume. We eliminated stocks that have discontinuities within this period (e.g. Facebook (FB) only started its trading mid 2012) and kept 94 large capital stocks. For each of the stock price time series, we calculate its return ratio R of stock prices time series $S(t)$ over a time interval Δt (1 day) as follows:

$$R(t) = (S(t) - S(t - \Delta t))/S(t - \Delta t) \quad (3)$$

Trading volume is the number of shares being transacted in a day. An example of the daily trading volume of AAPL, and the fraction of it in the portfolio throughout the year of 2013 (247 trading days) are shown in Figure 1.

3.2 Social Data

Wikipedia search volumes are obtained from the page view statistics [1] provided by Wikimedia dump service [2]. The data are saved every hour, containing request information of all *Wikipedia* pages in that hour. The information includes

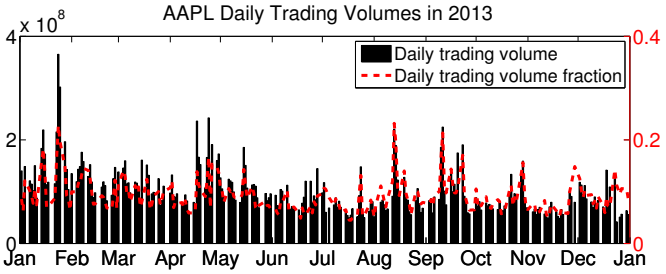

Figure 1: The daily trading volume and fraction of AAPL in 2013.

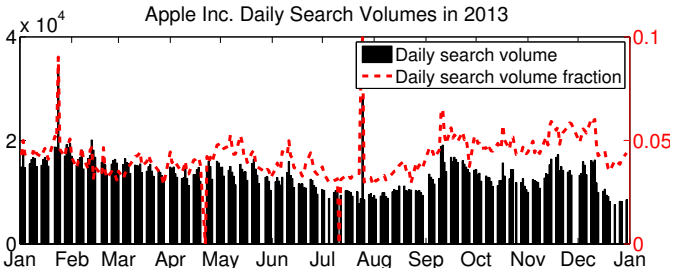

Figure 2: The daily search volume and fraction of Apple Inc. on Wikipedia in 2013.

the language of the page retrieved, the title of the page retrieved, the number of requests (further divided into the number of editing requests and reading requests), and the size of the content returned. The average size of the hourly file is 250 MB. The total data size of all pages requested throughout the year of 2013 is over 2.1 TB.

We gathered reading requests of the English pages of companies relevant to the top S&P 100 stocks we are interested in, and then aggregated the hourly volume into daily statistics for the same consistent 4 years. An example in Figure 2 shows the resulting daily search volume of Apple Inc., the company corresponding to the stock AAPL, throughout the year of 2013. Also shown is the search volume fraction normalized by the total search volume of all 94 companies.

4. METHODOLOGY

The general work flow is illustrated in Figure 3. We start with three types of data—the trading prices, trading volumes, and the corresponding *Wikipedia* query volumes of of a portfolio of stocks, and proceed in three steps. First, we attempt to find the equilibrium price for each stock in the portfolio. The residual of each of the stock is used as a trading signal indicator. When it is largely negative, the corresponding stock is to be longed, conversely when it is largely positive a stock is shorted. The threshold of residual value that triggers signal generation, as well as the number of components used are determined by training data (2010–2012) and applied to blind test data (2013).

Second, we investigate the hypothesis that the social behavior measured by *Wikipedia* query volume is predictive of stock trading volumes. Cross-correlation and Granger causality analyses are applied to the trading volume and corresponding *Wikipedia* query volume of the past n days. The outcomes of these analyses are, relatively, a correlation coefficient and a p-value.

In the third step, we take the residual-based trading signal, the trading volume (current) and and the *Wikipedia* query volume (current, but predictive to the future), and deploy enhanced trading strategies. Three types of strategies are designed, each leverages more information than its predecessor, with (1) residual signal only, (2) residual and Wiki, and (3) residual, Wiki and trading volume.

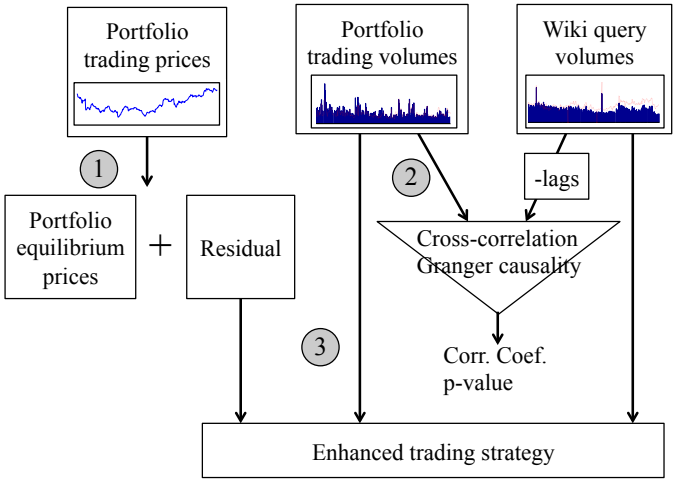

Figure 3: Diagram outlining 3 steps of methodology.

4.1 Does search volume cause trading volume?

We first associate each stock with its *Wikipedia* page. An alignment of dates is necessary since the stock is only traded on trading days (Monday to Friday excluding holidays) but page view data are gathered everyday. Lags are introduced so that the *Wikipedia* data is always n days prior to trading data, thus serving a predictive purpose.

Correlation and causality analyses are conducted at different time lags. The cross-correlation obtains a coefficient value R between two series A without shift and B shifted by a lag τ, along with the mean values of each. It is computed as:

$$R(\tau) = \frac{\sum_t \left((A(t) - \mu_A)\left(B(i-\tau) - \mu_B\right)\right)}{\sqrt{\sum_t \left(A(i) - \mu_A\right)^2}\sqrt{\sum_i \left(B(t-\tau) - \mu_B\right)^2}} \quad (4)$$

The Granger causality test [6] is performed to analyze causal relations. It is a statistical hypothesis test to determine whether A is useful in forecasting B, by building two autoregression processes for B, one without and the other with A as an external input, and examining the difference in accuracy with an F-test.

For both methods three years data of 2010–2012 (754 data points) are used. Figure 4 and Figure 5 show respectively the result of cross-correlation analysis and Granger causality test. Measurements are carried out on all 94 time series pairs at different time lags, and the distribution of values is seen in the form of box plots. The median value, 25th and 75th percentile as well as outliers are shown for each lagging condition. In each method besides the intended pair of search volume vs. trading volume (top), another pair of search volume vs. stock price (bottom) is also included for comparison. The return of price (Equation 3) is used in calculation. Results from both analyses show that the former pair has a stronger connection than the latter pair. In fact, search volume and stock price show no evident statistical correlation at all. As the lag increases the relationship weakens. In terms of the search volume vs. trading volume cross-correlation (Figure 4(a)), the median correlation of all stocks is highest when there is a 1-day lag between Wiki and trading. All lag conditions see positive correlation at the bottom 25th percentile and above. Granger causality outputs a p-value of a hypothesis test, the null hypothesis being the search volume do not Granger-cause the trading volume, or the stock price. From the result shown in Figure 5, the p-value is much lower in (a) compared to in that in (b). The null hypothesis is rejected so that we have a high level of confidence to say that the search volume Granger-causes trading volume. We observe that search volume has the highest Granger causality relation with trading volume for a lag of 1 day (p-value$<$0.05).

The rationale behind the observation that an increase of view counts of a company's Wiki page leads to an increase of future trading decisions, is that people tend to conduct "research" about companies prior to making decisions. The page viewing act represents the investor's behavior of gathering information and analyzing consequences. Note that the trading volume consists of both buying and selling actions. A stock is "hot" when the trading volume is high but not necessarily "in favor of".

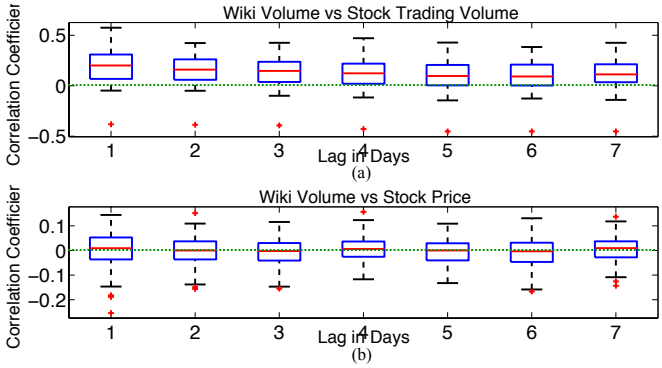

Figure 4: Correlation analysis: (a) between Wiki volume and stock trading volume, (b) between Wiki volume and stock price (return). Green dashed lines indicate the position of 0 correlation.

4.2 Portfolio trading strategy

We first introduce a baseline strategy and the basic form of group-wise pairs trading strategy.

- A percentile-based strategy. Everyday the best performing top 10 percentile and the worst performing 10 percentile stocks are selected. On the next day, go short the top percentile stocks and long the bottom.
- Portfolio-based pairs trading strategy. PCA is used to find the market fair value of each stock. Stocks are longed the next day if their current value is below their market fair value, and shorted otherwise.

We then propose two enhanced strategies that take into account social behavior trends. Each has two variations.

1. Enhanced strategy with residual + Wiki, denoted as ES-1. The residual is High if it is larger than zero, and

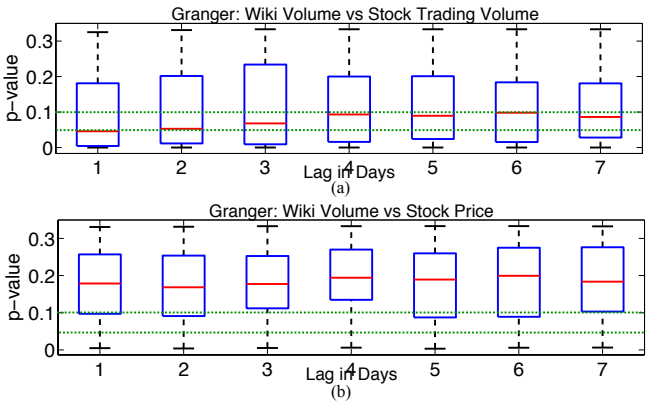

Figure 5: Granger causality: (a) between Wikipedia volume and stock volume, (b) between Wikipedia volume and stock price. Green dashed lines indicate the positions of $p = 0.1$ and $p = 0.05$.

is Low otherwise. The Wiki query volume is High if it is higher than its past 40-day moving average value, and is Low otherwise.

(a) If residual is Low, go long the stock; if residual is High and Wiki is High, go short the stock; otherwise do nothing.

(b) If residual is Low and Wiki is High, go long the stock; if residual is High and Wiki is High, go short the stock; otherwise do nothing.

2. Enhanced strategy with residual + Wiki + trading volume, denoted as ES-2. The High/Low criteria of residual and Wiki are the same as above. The trading volume is High if it is higher than its past 40-day moving average value, and is Low otherwise.

(a) If residual is Low, go long the stock; if residual is High and Wiki is High and trading volume is High, go short the stock; otherwise do nothing.

(b) If residual is Low and Wiki is High and trading volume is High, go long the stock; if residual is High and Wiki is High, go short the stock; otherwise do nothing.

This strategy is formed based on the assumption that whenever trading volume is too low, the stock is more prone to risk and its price movement tends to follow not evident trends but rather random walks. In this case we reduce the transactions conducted. We either close short positions (as in ES-1-a and ES-2-a) or be even more conservative and close both long and short positions (as in ES-1-b and ES-2-b).

5. BACK-TESTING RESULTS

The back-testing experiments consist in running the signals through historical data, and simulating trades according to designed strategies. We assume all trades are done at the closing price of that day, and in each transaction every stock is given the same amount of dollar value.

Daily Profit and Loss (PNL) can be calculated as the excess return gained divided by the investment amount. The cumulative PNL that totals all net profits and losses by successive addition of daily PNLs is then calculated. It results in a percentage of profit or loss over a period of time. The annualized Sharpe ratio [11] is also widely used by financial associations. The idea is to not only look at return amounts but also pay attention to risks. The ratio measures the excess return per unit of deviation.

$$Sharpe = \frac{\mu_{PNL(t)}}{\sigma_{PNL(t)}} * \sqrt{n} \quad (5)$$

where $\mu_{PNL(t)}$ and $\sigma_{PNL(t)}$ stand for the mean and standard deviation of daily PNL over some period of time. n is the number of trading days in a year (therefore annualized). We approximate $\sqrt{n}$ as 16. The Sharpe ratio gives higher values to better investment options.

First we compare two basic strategies. The 4-year-end PNLs and Sharpe ratios are shown in Table 1; the cumulative PNLs over the 4-year period are shown in Figure 6. The baseline percentile strategy is not performing well. Even though it ends at a 4-year-end PNL of 6.3585% by the end of 2013, along the time it easily got to a 0 return or even negative. Moreover, lots of oscillations are observed which indicate a high risk of this strategy. In contrary, the PCA based pairs trading is a much stabler strategy. The reason is that it attempts to find price discrepancy not only by looking at current return, but also turning to the history to find common components within a portfolio.

Table 1: Two Basic Strategies 4-Year Comparison

Strategy	4-year Sharpe Ratio	4-Year-end PNL
Baseline	0.2721	6.3585%
Pairs-PCA	1.5250	12.4633%

On top of PCA-based pairs trading, we proposed an enhanced strategy with four variations, depending on whether the trading volume is included and whether single or both sides of trades (long/short) are affected by Wiki activity information. Parameters are adjusted according to training data (01-04-2010 to 12-31-2012) and strategies are blind-tested on the 1 year test data between 01-02-2013 and 12-31-2013. Figure 7 shows the 1-year cumulative PNLs of them along with the two baselines. We can see that by including social activity information (all the ES-$*$), the performance is largely boosted. Adding trading volume into the formula does not seem to show much effect (comparing ES-1-$*$ with ES-2-$*$). Having social activity information impact short order decisions (removing short orders when Wiki volume is not as high) is better than having it impact both long and short orders (comparing ES-$*$-a with ES-$*$-b).

Table 2 displays the 2013 year-end PNL values and Sharpe ratios of all six strategies. While PNL might be the single number many individual investors care about, often the Sharpe ratio tells a better story. Comparing the first two rows, Baseline and Pairs-PCA, we found the PNLs are very close. However the Sharpe ratio is what tells these two strategies apart, at least from 2013's performance. The Baseline strategy was at negative gains for a lot of time during 2013, and even though it eventually picked up, the reliability of this strategy is far less compared to Pairs-PCA. Another interesting fact comes from a comparison of ES-1-b and ES-2-b. Although the former ended up with a slightly higher cumulative PNL, the latter had an advantage in Sharpe ratio. These two measures, when combined, are very helpful for picking up reliable strategies.

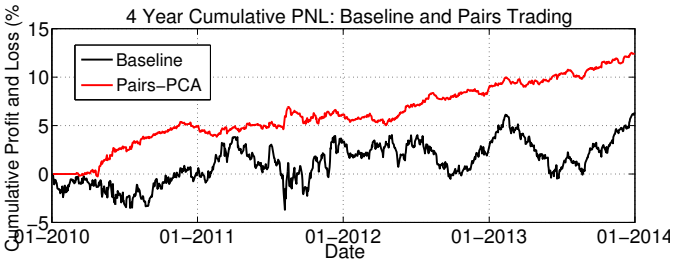

Figure 6: 4-year Cumulative PNLs of a percentile baseline strategy and the pairs trading with PCA.

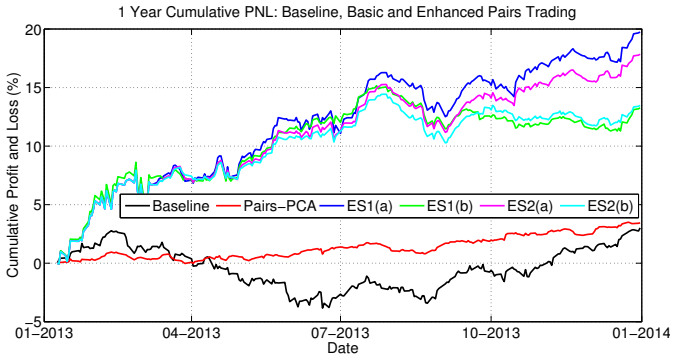

Figure 7: Cumulative PNLs of all strategies.

Table 2: Six Strategies 2013 Comparison

Strategy	1-year Sharpe Ratio	1-Year-end PNL
Baseline	0.6711	3.1113%
Pairs-PCA	2.0268	3.3697%
ES-1-a	2.6798	23.6996%
ES-1-b	2.0655	17.5623%
ES-2-a	2.6569	21.7989%
ES-2-b	2.2179	17.4297%

6. CONCLUSION

Investing in speculative assets such as stocks to generate profits is of both great interest and enormous challenge. A proper financial decision-making requires a profound understanding of the market, as well as people's behavior. Online social data from the increasingly popular social sites such as *Wikipedia* are an important resource for behavior study, which can help detect the trend of market movement and provide insights to trading strategy construction.

In this work, we formulate our strategy of profitable trading on the basis of two rationales: (1) the history tells a lot, and (2) the future can be indicated. We constructed a portfolio of top S&P 100 stocks, and leveraged their historical price data to reveal systematic components of a stock's movement, and regard their combined form as the relative market fair price that a stock is eventually going to converge to. A large and diverse portfolio is formed and PCA is used to separate the systematic components from idiosyncratic components. By doing this, the strategy outperforms an arbitrage baseline by over 6% in profit within a 4-year period.

Furthermore, we investigated the popular online encyclopedia *Wikipedia* page viewing data to examine the relationship of social activity and trading activity. We also studied whether such a relationship can be linked to subsequent decision making in the stock market. Our results are consistent with the hypothesis that *Wikipedia* data may have provided insights into future trends in the behavior of financial market actors. In our analysis, we find evident correlation and causality between the page view volume of articles relating to companies of stocks and the trading volume of that stock on a future day. Thus the social activity data can be regarded as an early indicator of market activity. Strategies are formed incorporating the indicator and profits of up to 23.7% are seen in the year of 2013. The same basic strategy without social modeling has achieved a profit of 3.37%. Thus over 20% of the gained profit comes from the information brought by social activity indicators.

Future work includes the investigation of semantic networks, public moods and social hotspots from more sophisticated and structured data (e.g. texts, sentiments). Data from various social sites such as Twitter, Google and blogs can be aggregated to provide different aspects of information regarding the stock trading.

7. REFERENCES

[1] Page view statistics http://dumps.wikimedia.org/other/pagecounts-raw/.

[2] Wikimedia downloads http://dumps.wikimedia.org/.

[3] Wikipedia http://www.wikipedia.org.

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

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

[6] C. W. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[7] H. Mao, S. Counts, and J. Bollen. Predicting financial markets: Comparing survey, news, twitter and search engine data. *arXiv preprint arXiv:1112.1051*, 2011.

[8] Y. Mao, W. Wei, and B. Wang. Twitter volume spikes: analysis and application in stock trading. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, page 4. ACM, 2013.

[9] H. S. Moat, C. Curme, A. Avakian, D. Y. Kenett, H. E. Stanley, and T. Preis. Quantifying wikipedia usage patterns before stock market moves. *Scientific reports*, 3, 2013.

[10] T. Preis, H. S. Moat, and H. E. Stanley. Quantifying trading behavior in financial markets using google trends. *Scientific reports*, 3, 2013.

[11] W. F. Sharpe. The sharpe ratio. *Streetwise–the Best of the Journal of Portfolio Management*, pages 169–185, 1998.

[12] T. B. Trafalis and H. Ince. Support vector machine for regression and applications to financial forecasting. In *Neural Networks, IEEE-INNS-ENNS International Joint Conference on*, volume 6, pages 6348–6348. IEEE Computer Society, 2000.

[13] G. Vidyamurthy. *Pairs Trading: quantitative methods and analysis*, volume 217. John Wiley & Sons, 2004.