THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

GOOGLE TRENDS PREDICT STOCK VOLATILITY

CHRISTOPHER SIERGIEJ
SPRING 2015

A thesis
submitted in partial fulfillment
of the requirements
for baccalaureate degrees
in Computer Science and Mathematics
with honors in Computer Science

Reviewed and approved* by the following:

Jesse Barlow
Professor of Computer Science
Thesis Supervisor

John Hannan
Professor of Computer Science
Honors Adviser

*Signatures are on file in the Schreyer Honors College.

Abstract

The thesis studies the effect of weekly search volume data from Google Trends on volatility measures of a portfolio of hand-picked stocks. Twelve stocks were selected from three sectors and a Granger causality analysis was performed to determine whether the search volume time series was useful in forecasting the volatility time series for a given stock. The results from the Granger causality analysis showed that some, but not all, stocks could use their search volume data from Google Trends to significantly forecast their volatility. For those stocks whose search volume data proved fruitful in forecasting their volatility, a search volume model consisting of lags of search volume data as predictors was compared to a null model consisting of the average of the volatility as a forecast. Using the mean absolute percentage error as a metric, the results support the view that the search volume model does have some forecast ability in producing volatility estimates.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction to the problem

As the internet becomes ubiquitous, the use of search engines are increasing. Naturally, interests arise to explore whether the search volume for a particular topic tells something about the collective sentiment of the topic for the general public. This research will explores the predictability of Google Trends on stock volatility. As Google holds a staggering 70% of U.S. search engine market share [1], the search volume from Google could be a reasonable measure of overall interest on the internet.

1.2 Background information

Stock market trading has high potential for data mining because of the enormous amount of historical data, which could suggest a competitive advantage over human inspection of the data. However, some researchers argue that the markets change so rapidly that there is no room to consistently obtain profits in their efficient markets hypothesis. According to the efficient markets hypothesis, stock prices are driven by new information rather than present and past prices. Because news is unpredictable, stock market prices will follow a random walk model and cannot be predicted by more than 50 percent accuracy.

Stock market prediction, the act of forecasting the future value of a stock or financial instrument traded on an exchange, is a problem that has been studied in both academia and industry. Clearly, a correct prediction could yield considerable returns for investors. However, some economists argue that the markets change so rapidly that it is impossible to forecast future value because current prices reflect all available information and that

stocks follow a random walk and thus are unpredictable in their Efficient-market hypothesis. The hypothesis has been empirically and theoretically disputed by investors and researchers who argue that cognitive biases such as overconfidence and overreaction are predictable.

There are three main types of prediction methods: fundamental analysis, technical analysis, and alternative methods. Fundamental analysis is concerned with looking at economic values and a company as a whole to measure the intrinsic value of a stock. On the other hand, technical analysis is not concerned with any fundamental company values but with potential trends from past market activity, such as past prices and volume. Rather than using a stocks intrinsic value, technical analysis uses charts to identify patterns and trends. Alternative methods such as using machine learning algorithms and time series analysis are utilized as well. While some of these methods could use technical analysis indicators, most of these methods include information beyond historical data about a stock.

1.3 Related research

Using Twitter data to forecast market movement: Several alternative methods are related to this research project. For example, several researchers have correlated sentiment from Tweets to the DJIA: Bollen, Mao, & Zeng used a Granger causality analysis and a self-organized fuzzy neural network to relate calm tweets to significantly predicting the DJIA in 2008 with 87.6% accuracy [2]. Additionally, Chyan, Hsieh & Lengerich used Twitter sentiment and neural networks to significantly predict the DIA [3]; Chakoumakos, Trusheim, & Yendluri used tweets containing company tickers, names, and nicknames and obtained a 78% accuracy to classify 10 tickers

in the NASDAQ [4]. Several other studies have been done finding Twitter data as a positive predictor by using support vector machines and logistic regression by Kuleshov [5]; Debbini, Estin & Goutagny [6]; Hsu, Shiu & Torczynski [7]; Mittal & Goel [8]; and Chen & Laze [9].

Using news articles to forecast market movements: Moreover, several researchers have correlated news articles to stock market price movements. Gidofalvi trained a naive Bayesian text classifier to predict if a stock would go up, down, or neither [10]. Lee and Timmons used a bag of words model on the New York Times articles from July 1994 to December 2004 to predict stock markets movements [11]. Finally, Schumaker and Chen used financial news to show that a support vector machine provides a statistically significant impact on predicting future stock prices compared to linear regression [12].

Using search volume words to forecast market movement: In addition, researchers have used search traffic on search engines to predict the stock market. Preis et al. described how certain words such as debt suggest that increases in search volume for financially relevant terms could predict losses in the stock market. Another project was done by Moat and Preis to link the number of views on Wikipedia articles on financial topics to large stock movements [13].

1.4 Goals

The overall goal of this thesis is to develop a prediction for future volatility that could be useful in volatility trading strategies. By looking at an overview of the stock market, to discussing the models that could be useful, and by discussing the results, the thesis could serve as a framework for

analyzing stock volatility.

1.5 Thesis structure

The thesis starts in chapter 2 with a discussion about related financial information, including options and their strategies, of which volatility affects. Next, chapter 3 examines the data acquisition process from Google Trends and the computation of several volatility measurements. After that, chapter 4 discusses the several models used to determine if search volume can significantly predict volatility and by how much. The thesis shows results in chapter 5 and a conclusion in chapter 6.

Chapter 2

Background information

2.1 Overview

This chapter starts with a discussion about the financial markets and trading. Then, the thesis explores options and their strategies for which volatility affects.

2.2 Financial theory

2.2.1 Markets

The financial markets are used by people to trade liquid assets at low transaction costs. The word liquid means that the asset can be easily transformed into cash. There are several types of markets from which one can trade. For example, the bond market and stock market allows financing through bonds and stocks. The commodity market facilitates trading of commodities. Money and insurance markets are used to provide short term debt financing and to redistribute various risks. Futures markets are a popular market that traders can trade on and provide forward contracts for trading products. Finally, derivatives markets provide contracts for the management of financial risk.

This thesis is one that attempts to forecast stock volatility using Google search volume data. Volatility plays an important role in derivatives contracts. Derivatives are contracts that derive their value from some underlying asset; examples include futures, forwards, and options. People who trade derivatives wish to gamble on the ratio of risk and return and on the asset to place their bet. Derivatives are used for speculation or to hedge risk.

2.2.2 Trading

This section will seek to describe the types of trading that could be profitable from this thesis. It first describes five kinds of trading: position trading, swing trading, intraday trading, high-frequency trading, and low-latency trading. These five types of trading occurs from the timescale of years and months to seconds and milliseconds. Obviously, as the speed of trade increases, the more likely computers will be the ones doing the trading.

Position trading is the longest type of trading and consists of holding a trading position for a few months. Fundamental analysis, or looking at companies values such as price-earnings ratio, are used to make decisions on these trades.

Next, swing trading is used for time periods of weeks and is used for more short-term fundamental analysis trading. Most of these traders attempt to speculate on price swings. This is an area for which volatility trading could be beneficial.

The next type of trading occurs daily, and that is intraday trading, where all positions will close before the end of the day. Both human and computer investors can do intraday trading.

The final two categories of trading are high-frequency trading and low-latency trading. Typically, these types of trades take on positions for milliseconds. Low-latency is described as a subset of high-frequency trading in that it has direct market access and attempt to arbitrage from price discrepancies in an order book.

2.3 Options

2.3.1 Overview

In finance, an option is a contact that allows the buyer, or owner of the option, the right, but not the obligation, to buy or sell an underlying asset at a specific price and time. Contrastingly, the seller has the obligation to perform the obligation. The premium is the cost to buy the right.

There are two kinds of options: a call option and a put option. The call option gives the buyer the right, but not the obligation, to buy an asset or underlying instrument. The put option gives the buyer the right, but not the obligation, to sell the asset or underlying instrument. Several specifications must be discussed prior to formation; those are listed in table 2.1.

Parameter	Description
Type	Call option or put option
Underlying asset	A stock; e.g. 500 shares of GOOG
Strike price	The price where the underlying asset will exercise
Expiration date	The date at which the contract expires
Settlement terms	Specifies whether the actual asset is delivered or cash

Table 2.1: Contract specifications of options.

2.3.2 Styles

There exist several styles of option. The most popular are European options and American options. The difference between the two is European options can only be exercised at expiry while American options can be exercised at or before expiry. More information about various option styles are shown in table 2.2.

Style	Description
European option	Can only be exercised at expiry.
American option	Can be exercised at or before expiry.
Bermudan option	Can be exercised at discrete periods at or before expiry.
Asian option	Payoff is determined by the average price over the lifetime of the option
Barrier option	Can only be exercised if the price meets a specific value, or ”barrier”
Binary option	Pays a certain amount only if the underlying asset meets a certain condition
Exotic option	Broad category that incorporates options with intricate structures
Vanilla option	Any option that is not exotic

Table 2.2: There exist several styles of options. Perhaps the most famous are European options and American options. The former only allows exercise at expiration, while the latter allows exercise at any time before or at expiration.

2.4 Strategies

This section will discuss several option trading strategies which volatility can be a factor. Of course, there are many option strategies that one could use; however, the discussion will concentrate on the four shown in table 2.3.

Strategy	**When to Use**	**Maximum Profit**	**Maximum Loss**	**Volatility**
Long Call	When the stock will go up in value	Unlimited	Cost of the contract	High volatility
Long Put	When the stock will go down in value	Strike price - cost of the contract	Cost of the contract	High volatility
Long straddle	When the stock will go up or down	Unlimited if the stock goes up, Strike price - cost of the contract if the stock goes down	Cost of the contract	High volatility
Long strangle	When the stock will go up or down but higher than that of a straddle	Unlimited if the stock goes up, Strike price - cost of the contract if the stock goes down	Cost of the contract	Very high volatility

Table 2.3: The table shows the strategies of options for a long call, long put, long straddle, long strangle, and a long put butterfly. The table lists when the buyer will want to run the option, the maximum potential profit and the maximum potential loss, and how volatility will affect the strategy.

2.4.1 Long call

A long call strategy gives the buyer the right to buy the underlying stock at a certain strike price. Compared to simply purchasing the stock, the buyer can profit if the stock rises, and can prevent the risk of the stock

falling in value that would result if the buyer owned the stock. The buyer of a long call will suspect that the stock will go up in value, and there is no maximum potential profit and the premium is the maximum potential loss. A way volatility could impact the strategy is that the buyer would hope that the stock will have high volatility, causing the price in go up. Figure 2.1 displays the profit of the long call.

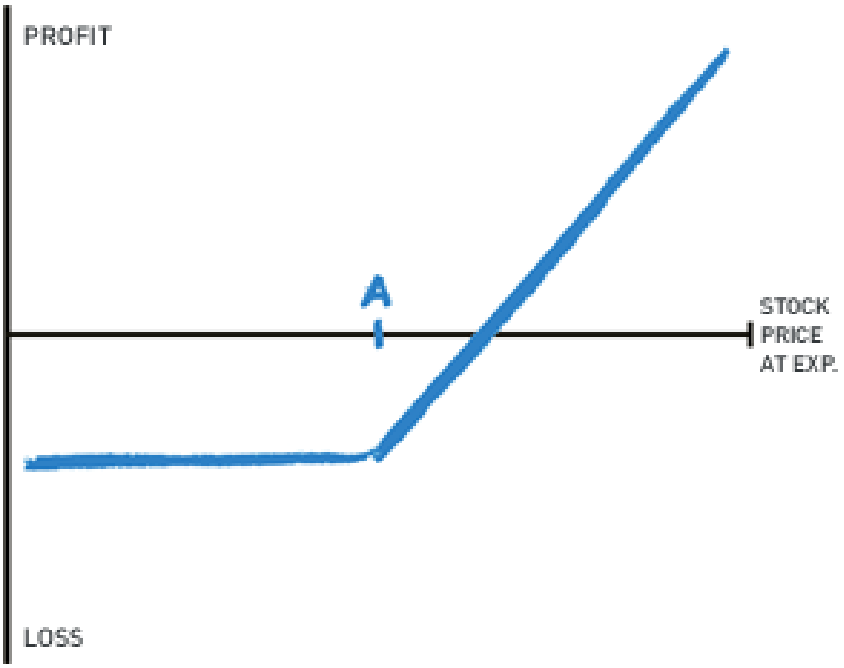

Figure 2.1: The profit of the long call. Notice how the there is no maximum potential profit and the premium is the maximum potential loss. Image obtained from Options Playbook.

2.4.2 Long put

A long put strategy gives the buyer the right to sell the underlying stock at a certain strike price. Compared to simply purchasing the stock, the buyer can profit if the stock falls. The buyer of a long put will suspect that the stock will go down in value, and there is a maximum potential profit of the strike price minus the price of the contract, and a maximum potential profit of the price of the contract. Like a long call, the buyer could

hope for period of high volatility such that the stock will decrease in value. Figure 2.2 displays the profit of the long put.

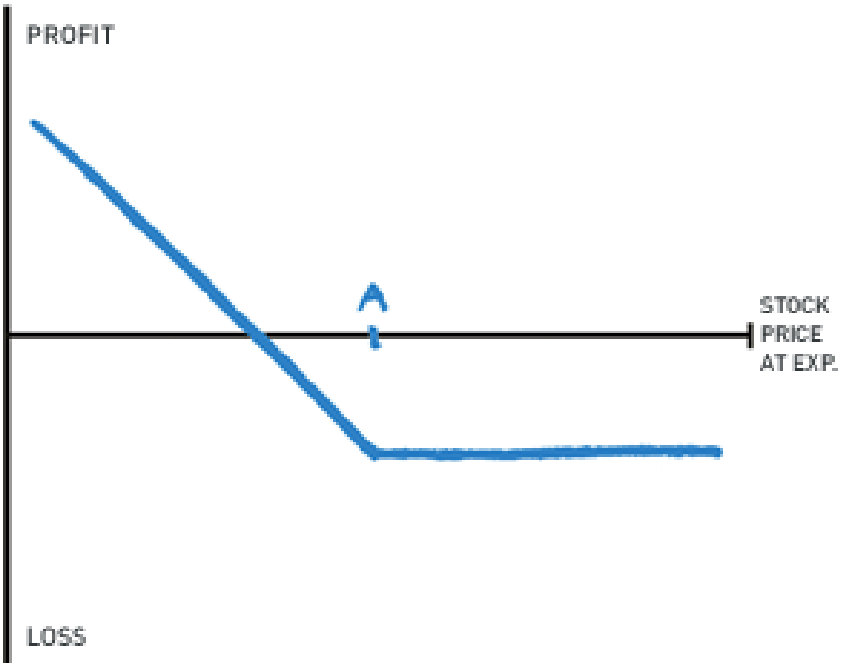

Figure 2.2: The profit of the long put. Notice how the there is a maximum potential profit of the strike price minus the price of the contract, and a maximum potential profit of the price of the contract. Image obtained from Options Playbook.

2.4.3 Long straddle

A long straddle gives the best of the long call and long put, giving you the right to buy and the right to sell. But because of the increased opportunity, the cost for this contract is greater than that of the long call or the long put. Traders who use a long straddle will seek to profit from the stock changing price in either the up or down direction. The buyer should use a long straddle when the buyer suspects periods of high volatility and has no idea which direction the stock will go. The maximum potential profit is unlimited if the stock goes up and the strike price minus the price of the contract if the stock goes down. The maximum potential loss is the price

of the contract. Figure 2.3 displays the profit of the long straddle.

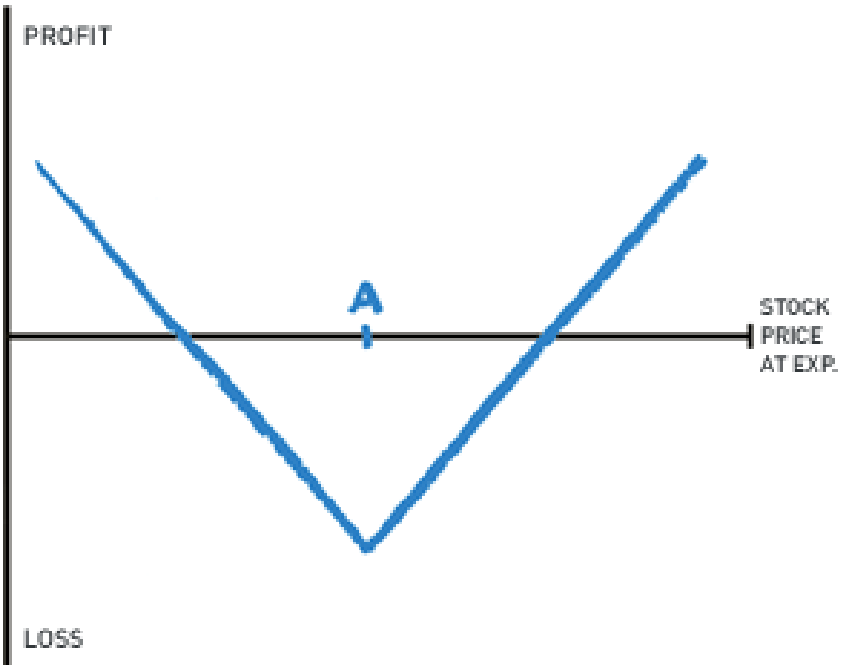

Figure 2.3: The profit of the long strangle. Notice how the the maximum potential profit is unlimited if the stock goes up and the strike price minus the price of the contract if the stock goes down, and the maximum potential loss is the price of the contracts. Image obtained from Options Playbook.

2.4.4 Long strangle

A long strangle is like a long straddle in that there is a hope that the stock will be volatile. However, the strangle requires a significant price change compared to the straddle even to be profitible. The strangle consists of buying a long call and a long put with different strike prices, often out of the money, which reduces the contract prices compared to the straddle. The maximum potential profit is unlimited if the stock goes up and the strike price minus the price of the contract if the stock goes down. The maximum potential loss is the price of the contract. Figure 2.4 displays the profit of the long strangle.

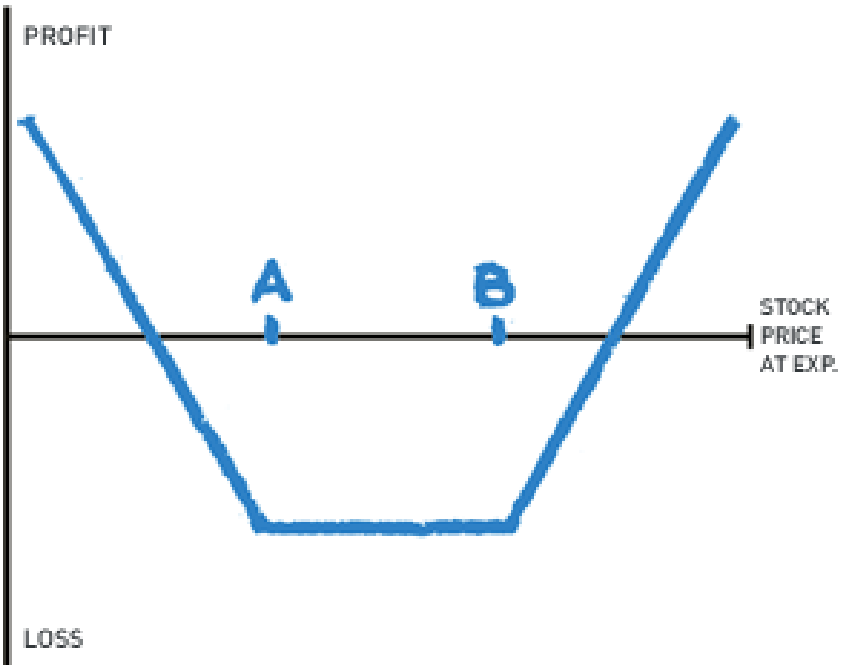

Figure 2.4: The profit of the long strangle. Notice how the the maximum potential profit is unlimited if the stock goes up and the strike price minus the price of the contract if the stock goes down, and the maximum potential loss is the price of the contracts. Image obtained from Options Playbook.

Chapter 3

Data

3.1 Overview

The data chapter starts with a discussion about the extraction process for the Google Trends search volume data. Then, the thesis considers various volatility metrics and their strengths and weaknesses.

3.2 Acquisition

The search volume data for all symbols of the stocks in table 3.1 is extracted from Google Trends. It is aggregated weekly from January 1, 2004 to December 31, 2014. The search volume is standardized such that the week with the highest search volume is 100 and the week with the lowest search volume is 0. Thus, two terms cannot compare. A time series for the search volume for Apple Inc. (search term AAPL) can be seen in Figure 3.1.

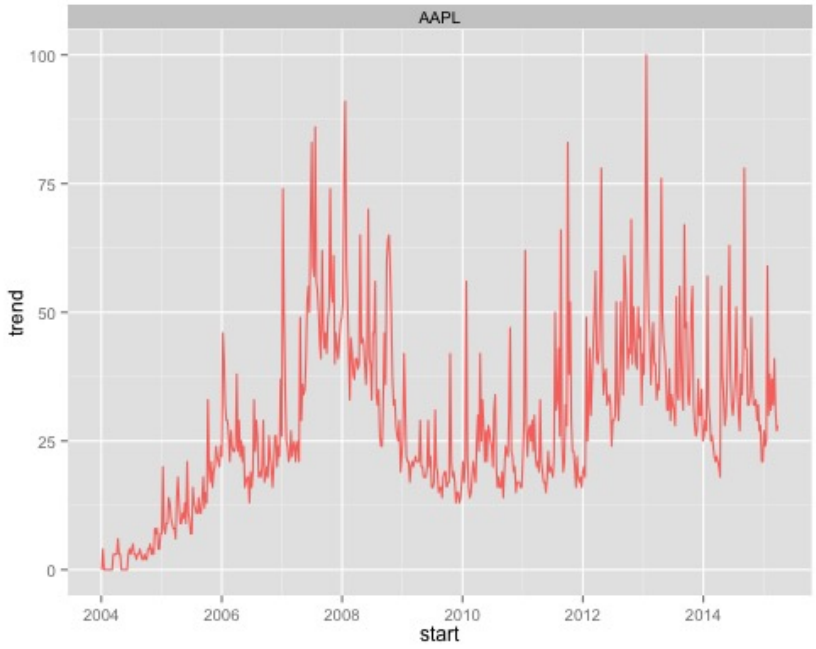

Figure 3.1: Time series for search volume of Apple, Inc. given by its ticker AAPL. Notice how the minimum and maximum values are 0 and 100 respectively.

The daily open, high, low, close, adjusted close, and volume prices are extracted from Yahoo! Finance for the same period as the search volume data. The stocks selected are a predetermined list of stocks from various sectors and are high in popularity as one could expect more data available. Furthermore, the stocks selected contain a unique symbol that most likely will not cause noise in the search volume data. For example, the term IPA could refer to the pharmaceutical company, Interpharm Holdings Inc, or an India Pale Ale. The stocks selected are shown in Table 3.1.

Sector	Stock	Symbol
Technology	Apple Inc.	AAPL
	Microsoft Corporation	MSFT
	Yahoo! Inc.	YHOO
	NVIDIA Corporation	NVDA
	Applied Materials Inc.	AMAT
	Hewlett-Packard Company	HPQ
Financial	JPMorgan Chase	JPM
	Wells Fargo & Co	WFC
	Bank of America Corp	BAC
Industrial	Lincoln Electric Holdings, Inc.	LECO
	AGCO Corporation	AGCO
	Manitowoc Company Inc	MTW

Table 3.1: A listing of stocks selected for this study including their sector, company name, and ticker symbol.

3.3 Feature construction

Once the stock market data is obtained, volatility can be computed. Volatility is the measure for variation of a price of a financial instrument over time. There are two kinds of volatility: historical volatility and implied volatility. The former is computed using historical time series data on a stock while the latter is computed primarily on options using an options pricing model and returning a theoretical value equal to the current price

of the option. This research considers historical volatility only.

There exist several methods of measurements for historical volatility including close-to-close volatility, Parkinson volatility, Garman-Klass volatility, Rogers-Satchell volatility, and Yang-Zhang volatility. Brief explanations and formulas are given below.

A closed form formula for close-to-close volatility is given in Equation 3.1, where N is the number of periods per year, n is the number of periods for the volatility estimate, and c_i is the closing price for time i.

$$\sigma_{cc} = \sqrt{\frac{N}{n-1}\sum_{i=1}^{n} \ln \frac{c_i}{c_{i-1}}} \tag{3.1}$$

The close-to-close volatility measure is the more simplistic measures that only uses closing prices to determine its value. On the other hand, the Parkinson volatility estimator given in equation 3.2 uses high and low prices. Let N be the number of periods per year, n be the number of periods for the volatility estimate, and h_i and l_i be the high and low prices, respectively, for time i.

$$\sigma_p = \sqrt{\frac{N}{n}\frac{1}{4\ln 2}\sum_{i=1}^{n}\left(\ln \frac{h_i}{l_i}\right)^2} \tag{3.2}$$

One drawback for the Parkinson volatility estimator is that it underestimates volatility because it assumes continuous trading. An extension of the Parkinson volatility estimate is the Garman-Klass volatility estimator found in equation 3.3. In addition to using the high and low prices, it uses open and close prices. If N is the number of periods per year, n is the number of periods for the volatility estimate, and h_i, l_i, o_i, c_i are the high, low, open, and close prices, respectively, for time i, then the Parksin volatility

estimate is:

$$\sigma_{gk} = \sqrt{\frac{N}{n}\frac{1}{2}\sum_{i=1}^{n}\left(\ln\frac{h_i}{l_i}\right)^2 - \left(2\ln 2 - 1\right)\left(\ln\frac{c_i}{o_i}\right)^2} \tag{3.3}$$

While the Garman-Klass volatility estimator uses the most information of the available data, it underestimates volatility by ignoring overnight jumps. The Rogers-Satchell volatility estimator in equation 3.4 is different than the previous three volatility estimators in that it handles a drift, or a non-zero mean. Let N be the number of periods per year, n be the number of periods for the volatility estimate, and h_i, l_i, o_i, c_i are the high, low, open, and close prices, respectively, for time i.

$$\sigma_{rs} = \sqrt{\frac{N}{n}\sum_{i=1}^{n}\ln\frac{h_i}{c_i}\ln\frac{h_i}{o_i} + \ln\frac{l_i}{c_i}\ln l_i o_i} \tag{3.4}$$

The Rogers-Satchell volatility estimator is good in that it allows for the presence of trends; however, it still cannot deal with jumps. Thurs, the Yang-Zhang volatility estimator in equation 3.5a resolves these issues. Let N be the number of periods per year, n be the number of periods for the volatility estimate, and h_i, l_i, o_i, c_i are the high, low, open, and close prices, respectively, for time i.

$$\sigma = \sqrt{N}\sqrt{\sigma_o^2 + k\sigma_c^2 + (1-k)\sigma_{rs}^2} \tag{3.5a}$$

$$\sigma_o^2 = \frac{1}{n-1}\sum_{i=1}^{n}\ln\left(\frac{o_i}{c_{i-1}}\right)^2 \tag{3.5b}$$

$$\sigma_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}\ln\left(\frac{c_i}{o_{i-1}}\right)^2 \tag{3.5c}$$

$$\sigma_{rs}^2 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} \ln\frac{h_i}{c_i}\ln\frac{h_i}{o_i} + \ln\frac{l_i}{c_i}\ln l_i o_i} \tag{3.5d}$$

$$k = \frac{0.34}{1+\frac{n+1}{n-1}} \tag{3.5e}$$

The Yang-Zhang volatility estimator is specifically designed to have minimum estimation error in that it can handle both drift and jumps. However, the performance degrades when the process is dominated by jumps.

A summary of the five volatility measures can be found in table 3.2.

Estimate	Prices	Handle drift	Handle jumps
Close-to-close	Close	No	No
Parkinson	High, Low	No	No
Garman-Klass	Open, High, Low, Close	No	No
Rogers-Satchell	Open, High, Low, Close	Yes	No
Yang-Zhang	Open, High, Low, Close	Yes	Yes

Table 3.2: A listing of volatility estimates and the prices used to compute them. Also indicates whether the volatility estimates can handle drift and jumps.

Chapter 4

Models

4.1 Overview

This chapter discusses the various statistical models. First, the thesis examines the Granger causality test which is used to determine whether search volume can significantly predict volatility. Then, a discussion about regression models including the null regression model, simple linear regression model, and support vector regression model takes place. Finally, the chapter ends with an examination about metrics and cross-validation procedures.

4.2 Granger causality

The Granger causality test is a statistical hypothesis test to determine if one time series is useful in forecasting another. Originally proposed by Clive Granger, British economist who was awarded the Nobel Memorial Prize in Economic Sciences, the Granger causality test is a measure if the future values of a time series can be determined using past values of another time series. Contrasting with regression, Granger argued that regression reflects correlation while the causality test finds predictive causality. Therefore, the Granger causality test goes further than just calculating correlation between two variables and calculates the likelihood of a true causal relationship.

Granger defined causality as the following: the cause happens prior to its effect and the cause has unique information about future values of its effects.

The volatility time series, denoted V_t, is defined to reflect weekly volatility for a given stock, i.e. its values are the historical volatility between

weeks t and $t-d$. To test whether our Google Trend search volume time series predict changes in stock volatility we compare the variance explained by the two linear models shown in Equation 4.1 and Equation 4.2. The first model L_1 uses only n lagged values of V_t, i.e. $D_{t-1}, D_{t-2}, ..., D_{t-n}$ for prediction, while the second model L_2 uses the n lagged values of both V_t and the Google Trend search volume time series denoted $T_{t-1}, T_{t-2}, ..., T_{t-n}$.

$$L_1 : V_t = \alpha + \sum_{i=1}^{n} \beta_i V_{t-i} + \epsilon_t \tag{4.1}$$

$$L_2 : T_t = \alpha + \sum_{i=1}^{n} \beta_i V_{t-i} + \sum_{i=1}^{n} \gamma_i X_{t-i} + \epsilon_t \tag{4.2}$$

Seasonality and trend are checked for all time series data using various applicable tests. Results can be found in the next section.

4.3 Regression

Regression is used to forecast values given a set of input data. There are multitudes of regression techniques; however, for this research the models being considered are the null regression model, simple linear regression, and a support vector regression. See table 4.1 for more information.

Model	Independent Variables
Null regression model	Average value of volatility estimate
Linear regression	Lags of search volume Lags of volatility Lags of search volume and volatility
Support vector regression	Lags of search volume Lags of volatility Lags of search volume and volatility

Table 4.1: A table listing of the three main models this research will use. The first is the null regression model consisting of the average of the volatility as a predictor. The next are three linear regression models consisting of lags of search volume, lags of volatility, and lags of search volume and volatility. Finally, three support vector models are used consisting of lags of search volume, lags of volatility, and lags of search volume and volatility.

4.3.1 Null regression model

In order to have a benchmark regression model to compare others against, a null regression model is used. The null regression model could be thought of as the "obvious guess" that other models must do better than. The null model in this situation is the simplest possible value. In the case of this research project, it will be the mean value of the volatility output. It could be thought of a lower bound on model performance that other models should perform better than.

Let y_i consist of the volatility at time i, and let μ be a function that computes the average of all volatility values, and let y consist of the vector of all values. The null regression model is defined in equation 4.3 and equation 4.4.

$$y_i = \mu(y) \tag{4.3}$$

$$\mu(y) = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{4.4}$$

4.3.2 Simple linear regression

Simple linear regression is a model to estimate a dependent variable with independent variables. It attempts to fit a straight line through a set of n points in such a way that minimizes the sum of squared residuals, or the distances between the predicted dependent variable value and the actual dependent variable value.

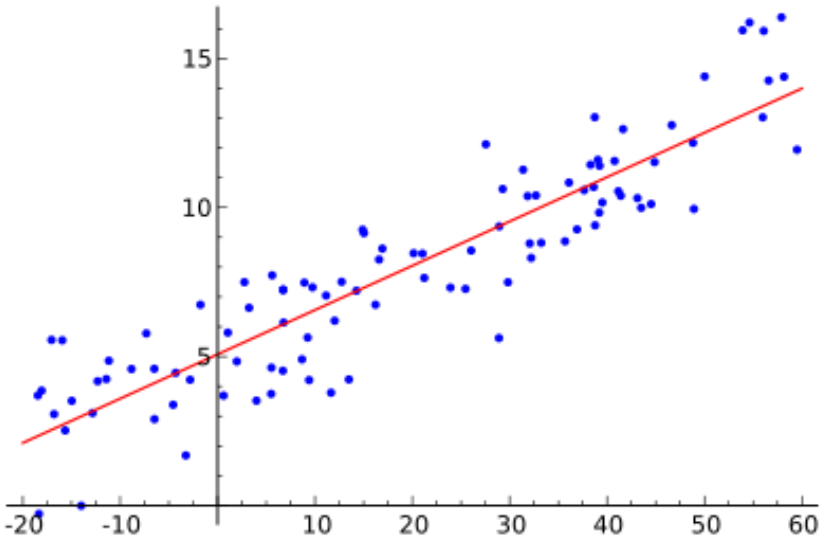

Figure 4.1: Example of linear regression on data. The plots consist of data points representing observations. The line is the fitted regression line for the data. Image obtained from Wikipedia.

Let y_i be the volatility measure at time i, and let x_i be the search volume measure at time i. The lags of the volatility measure y_{i-1} and y_{i-2} correspond to the previous observation and the observation before the previous observation. Likewise, x_{i-1}, x_{i-2}, x_{i-3}, and x_{i-4} are the four previous values of the search volume for time i.

There are three simple linear regression models being considered. The first contains only search volume as predictors in equation 4.5, the second contains only volatility as predictors in equation 4.6, and the third contains both search volume and volatility as predictors in equation 4.7.

$$y_i = \beta_1 x_{i-1} + \beta_2 x_{i-2} + \beta_3 x_{i-3} + \beta_4 x_{i-4} + \epsilon_i \tag{4.5}$$

$$y_i = \beta_1 y_{i-1} + \beta_2 y_{i-2} + \epsilon_i \tag{4.6}$$

$$y_i = \beta_1 x_{i-1} + \beta_2 x_{i-2} + \beta_3 x_{i-3} + \beta_4 x_{i-4} + \beta_5 y_{i-1} + \beta_6 y_{i-2} + \epsilon_i \tag{4.7}$$

Where y_i is the volatility estimate for time i, x_i is the search volume estimate for time i, β_i is the parameter vector or regression coefficient, and ϵ_i is the error term.

Linear regression has several assumptions. The first is linearity or that the mean of the response variable is the linear combination of the parameters. Next is that the errors have a constant variance and do not exhibit heteroscedastic behavior, or when a variable's variance is unequal across the values of another variable that serves as a predictor. Then, the errors must be independent and uncorrelated with each other. And finally, that the predictors do not exhibit multicollinearity, or when two or more predictors are highly correlated. Several of these assumptions can be tested through statistical tests and through inspection of graphs.

4.3.3 Support vector regression

A support vector machine is a supervised machine learning that can be used to perform regression analysis. The support vector regression defined in equation 4.8 and equation 4.9 seeks to solve the following.

$$\min \frac{1}{2}||w||_2 \tag{4.8}$$

$$\text{such that} \begin{cases} y_i - \langle w, x_i \rangle - b \leq \epsilon \\ \langle w, x_i \rangle + b - y_i \leq \epsilon \end{cases} \tag{4.9}$$

In the equations, x_i represents the predictors, y_i represents the dependent variable, ϵ is the threshold such that all predictions have to be within ϵ range of the true value. See figure 4.2 for more information.

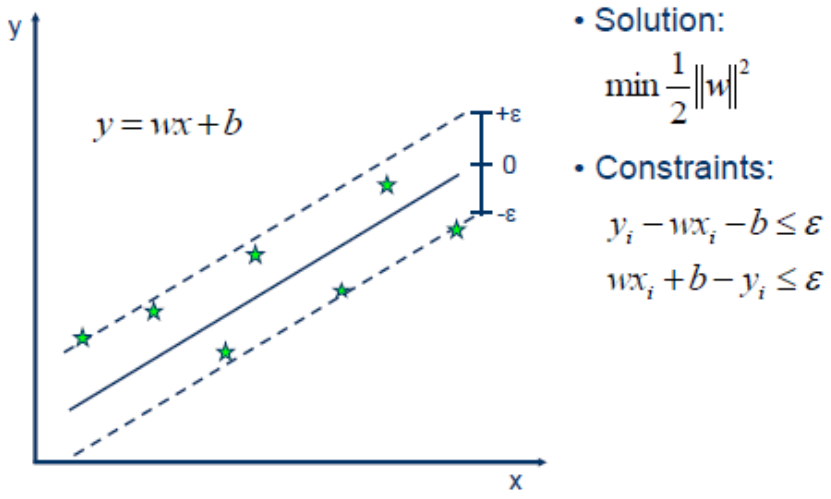

Figure 4.2: Example of support vector regression on data. Image obtained from Dr. Saed Sayad.

Similar to the linear regression models, this research will use three support vector regression models consisting of only search volume as independent variables, only volatility as independent variables, and both search

volume and volatility as independent variables.

Support vector regression are less restrictive than that of linear regression. Additionally, support vector machines tend to not over fit the data. This will serve as a comparison to the model produced by linear regression.

4.4 Miscellaneous matters

4.4.1 Metrics

In order to evaluate our models, several metrics will be used. They are the mean absolute error, mean squared error, root mean squared error, and the mean absolute prediction error.

The mean absolute error is used to compare forecasts to its outcome. It is defined in equation 4.10. Let n be the total number of observations, $\hat{Y}_i$ be the predicted value, and Y_i be the actual value.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{Y}_i - Y_i| \tag{4.10}$$

It is the average of the absolute errors. The next statistics is the mean squared error given in equation 4.11. It measures the average of the squares of the errors. Let n be the total number of observations, $\hat{Y}_i$ be the predicted value, and Y_i be the actual value.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 \tag{4.11}$$

The next statistic is the root-mean-squared error. It is defined in equa-

tion 4.12.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|\hat{Y}_i - Y_i|} \tag{4.12}$$

The last statistic is probably the most used statistic to evaluate accuracy of a time series in trend estimation. It is called the mean absolute percentage error and is defined by the formula in 4.13.

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{\hat{Y}_i}\right| \tag{4.13}$$

4.4.2 Cross-validation

In order to benchmark our performance, two models will be used with different predictors. For each model, the dependent variable is stock volatility. For the first model, the independent variables are lags of the search volume, and for the second model, the independent variable is the lag of the stock volatility.

The second model is known as the naive approach to forecasting that is used as a benchmark to use against more sophisticated models. Because the first differences of the time series show that the data is stationary, the naive forecast equals the previous period's value.

Both models will be fitted using simple linear regression and a support vector regression with a linear kernel.

Cross-validation is a statistical technique for determining how a given model will generalize to an independent data set. Generally, cross-validation is used when one wants to estimate how accurate a prediction will perform in practice. From all the data, there will exist a subset of data consisting of a training set and a testing set. The training set consists of the known data

from which the model is fitted and run, and then the testing set consist of the unknown data against which the model is tested. Cross-validation uses a validation data set, or a subset of the training data set, that acts as the test set in the model building phase. The goal is to reduce over-fitting and see how well the model will be applicable to data sets it has not seen. Multiple rounds of cross-validation are used with different partitions and the validation results are averaged to reduce variability of model statistics.

General cross-validation procedures include exhaustive cross validation, where learning and validating takes place on all possible partitions of a data set, and non-exhaustive cross validation, which do not use all possible partitions. A popular non-exhaustive cross validation technique is k-fold cross validation in which the training set is divided into k partitions and the model is fitted on the $k-1$ partitions and is validated on the remaining partition.

However, the data in the problem exhibits an ordering because they are associated with a specific time. Therefore, general cross-validation approaches will not work. For example, consider a cross-validation approach in our problem that trains a model with data from year 2010 and validates the model with data from year 2005. Clearly, it does not make sense to predict values from the past with values from the future.

In order to deal with the issue of time ordered data, three solutions are considered. The first is to use a one shot testing approach in which the training data is split into a training and testing set. A positive for this approach is that it will be computationally less intensive as there will be one fitting procedure and one testing procedure. A negative for this approach is that the statistics produced from the model, for example, the estimated mean absolute prediction error, may be biased because there is only one

value.

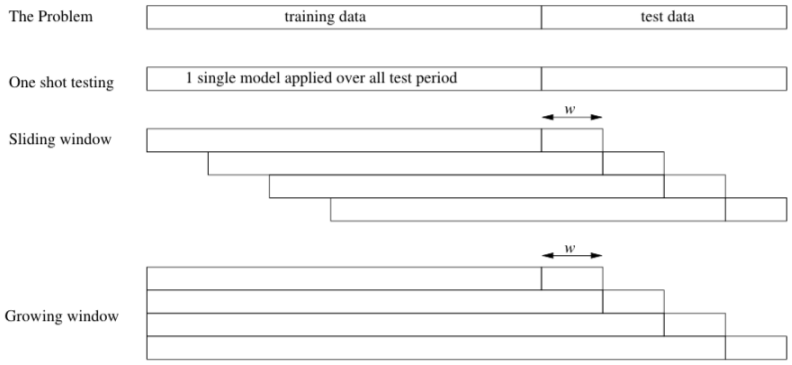

Figure 4.3: Examples of different time series cross validation methods. The three methods are one shot testing, a sliding window, and a growing window. Image obtained from Luis Torgo.

The next approach is to use a growing window technique. The growing window will add values into the training set and validate on a constant sized validation set. So, the growing window approach may train on the first five values and validate on the sixth, and then train on the first six values and validate on the seventh, and so on. A positive is that the statistical estimates will be more accurate than the one shot testing approach. A negative is that the growing window may not be able to account for shifts in the time series because it will train on all past values. A sliding window approach is discussed to overcome this problem.

The sliding window approach consists of training on n observations and validating on m observations and shifting the training and shifting the sets. For example, a sliding window approach may train on the first to fifth values and validate using the sixth, and then the next sliding window will train on the second to sixth and validate using the seventh.

To summarize, there exist three techniques to apply cross-validation to ordered time series data: one shot testing, growing window, and the sliding window approach. For the sake of this research, the sliding window

approach will be used.

Chapter 5

Results

5.1 Overview

The results chapter discusses the outcomes of the models in the previous chapter. First, the Granger causality test shows that search volume is significant in predicting volatility for some stocks. Then, the regression models show that search volume performs better than the null regression model.

5.2 Granger causality

5.2.1 Overview

The research seeks to answer if Google search volume can significantly forecast stock volatility. Consider testing to see if Microsoft stock volatility is caused by Google Trends search volume. The plot in figure 5.2 of the Yang Zhang volatility estimate with $n = 3$ weeks for 2014 Microsoft stock is shown in red below, and the 2014 MSFT Google Trend data is shown in blue below. In order to plot the two time series together, the volatility time series and search volume time series were scaled and centered by equation 5.1 where μ is the mean function and σ is the standard deviation function.

$$Z_i = \frac{X_i - \mu(X)}{\sigma(X)} \tag{5.1}$$

The following is a plot in figure 5.3 of the search volume time series which may indicate that it is not stationary. The Kwiatkowski Phillips Schmidt Shin (KPSS) test is a statistical test to determine whether the time series is stationary. The KPSS test was ran to find the least number of differences

required to fail the test at a significance level of 0.05. Results indicate that the series should be difference by 1. The series was also checked for seasonality using the Canova-Hansen test using a similar procedure to find the necessary amount of differencing.

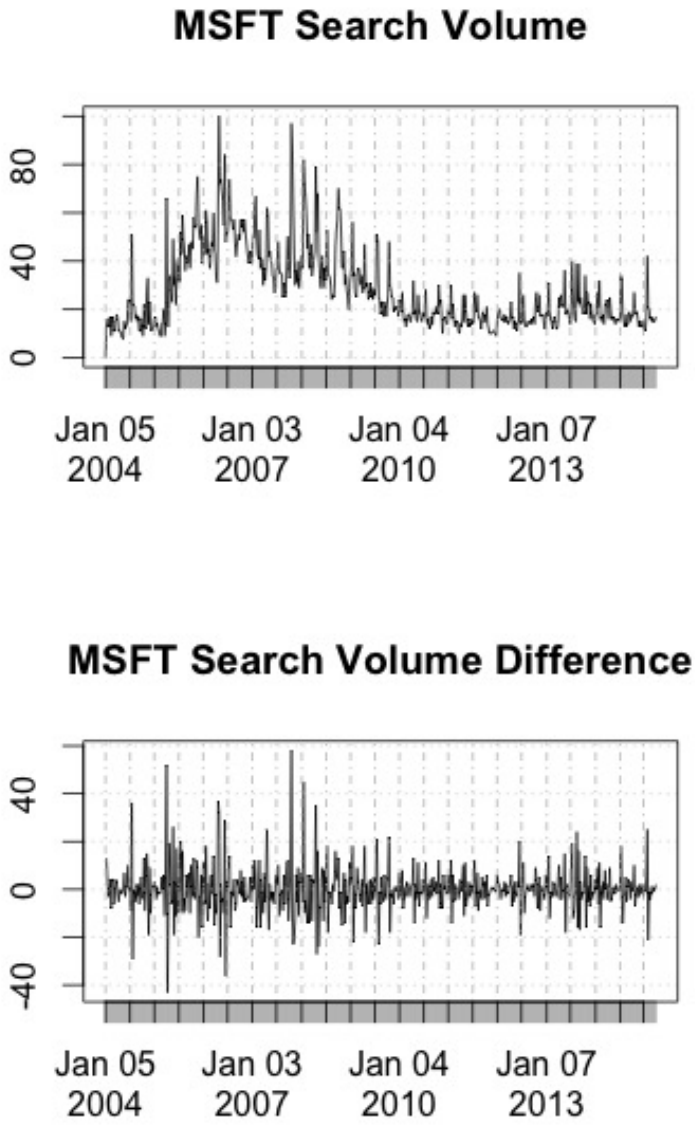

Figure 5.1: The plot contains two graphs. The first graph is a time series plot for the Microsoft search volume which looks like it may not be stationary. After running the KPSS test, the time series is differenced by one. The second plot is the time series for the Microsoft search volume after being difference by one.

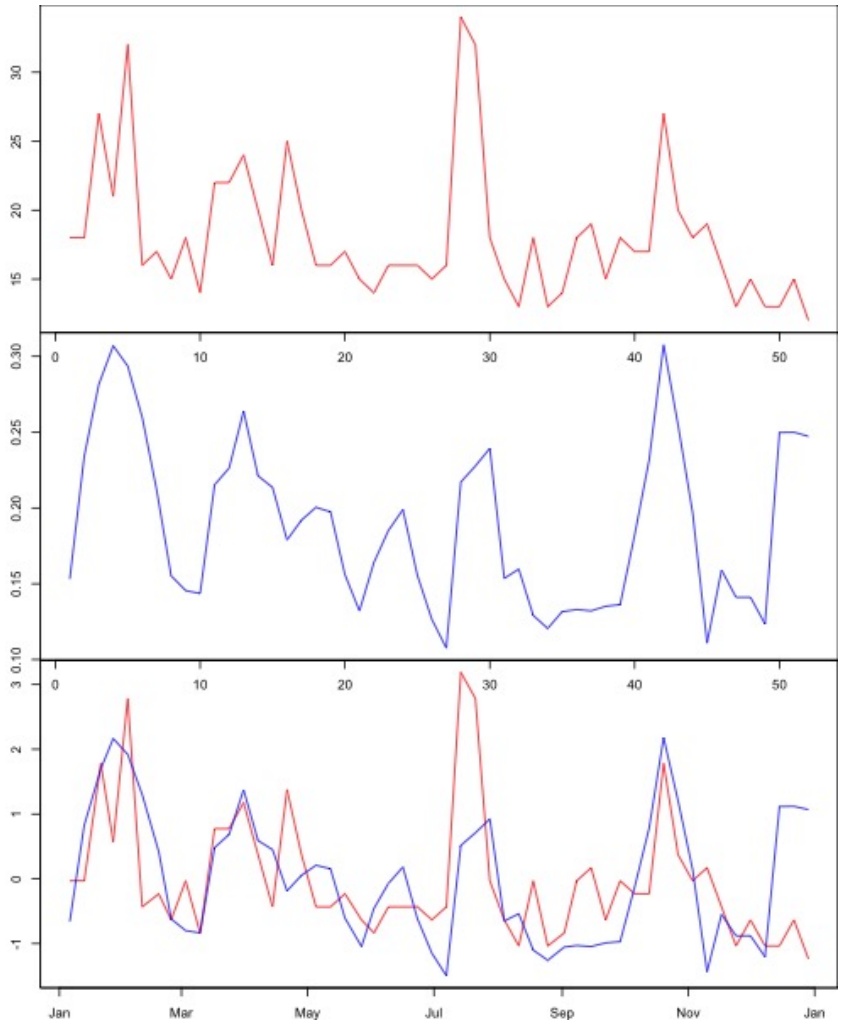

Figure 5.2: The plot contains three graphs. The first graph is the time series of the 2014 Google Trend data of the search volume shown in red. The second graph is the time series of the 2014 Yang Zhang volatility estimate with $n = 3$ for the Microsoft stock shown in blue. Finally, centered and scaled versions of the volatility and search volume time series are plotted together in the third graph.

After making the time series stationary and adjusting for seasonality, the Granger causality tests are run for the Microsoft search volume and the Yang-Zhang volatility estimator for $n \in (3, 4, 5)$ number of weeks to use in the volatility calculation and $l \in (1, 2, 3, 4)$ number of lagged weeks to use in the model. The null and alternative hypotheses are H_0 : Google Trends search volume data for Microsoft does not cause Yang-Zhang volatility for

Microsoft versus H_A : Google Trends search volume data for Microsoft does cause Yang-Zhang volatility for Microsoft. Results from the statistical experiments are in table 5.1.

Number of weeks for volatility calculation	Number of lagged weeks in model	P-value
2	1	7.553e-08 *
2	2	1.549e-05 *
2	3	5.112e-06 *
2	4	8.153e-06 *
3	1	0.06459
3	2	1.003e-11 *
3	3	2.182e-08 *
3	4	7.358e-08 *
4	1	0.2415
4	2	0.004794 *
4	3	1.472e-11 *
4	4	6.093e-06 *

Table 5.1: A table of Granger causality analysis test results for Microsoft search volume and Yang-Zhang volatility estimate. The first column contains the number of weeks for computing the Yang-Zhang volatility estimate and ranges from 2 to 4 weeks. The second column corresponding to the number of lagged terms used in the Granger causality model. The third column corresponds to the p-value of statistical test where the asterisk symbol is used for tests with significant results.

The results from the statistical test show that Microsoft search volume data significantly causes Microsoft Yang-Zhang volatility.

Suppose there is interest in determining whether the Microsoft Yang-Zhang volatility causes the Google Trends search volume for Microsoft. The results for the statistical tests are shown in table 5.2. The results show that most of the tests are not statistically significant.

Number of weeks for volatility calculation	Number of lagged weeks in model	P-value
2	1	0.09842
2	2	0.09925
2	3	0.07984
2	4	0.03042 *
3	1	0.09842
3	2	0.09925
3	3	0.07984
3	4	0.03042 *
4	1	0.09842
4	2	0.09925
4	3	0.07984
4	4	0.03042 *

Table 5.2: A table of Granger causality analysis test results for Yang-Zhang volatility estimate causing Microsoft search volume. The first column contains the number of weeks for computing the Yang-Zhang volatility estimate and ranges from 2 to 4 weeks. The second column corresponding to the number of lagged terms used in the Granger causality model. The third column corresponds to the p-value of statistical test where the asterisk symbol is used for tests with significant results.

5.2.2 Summary

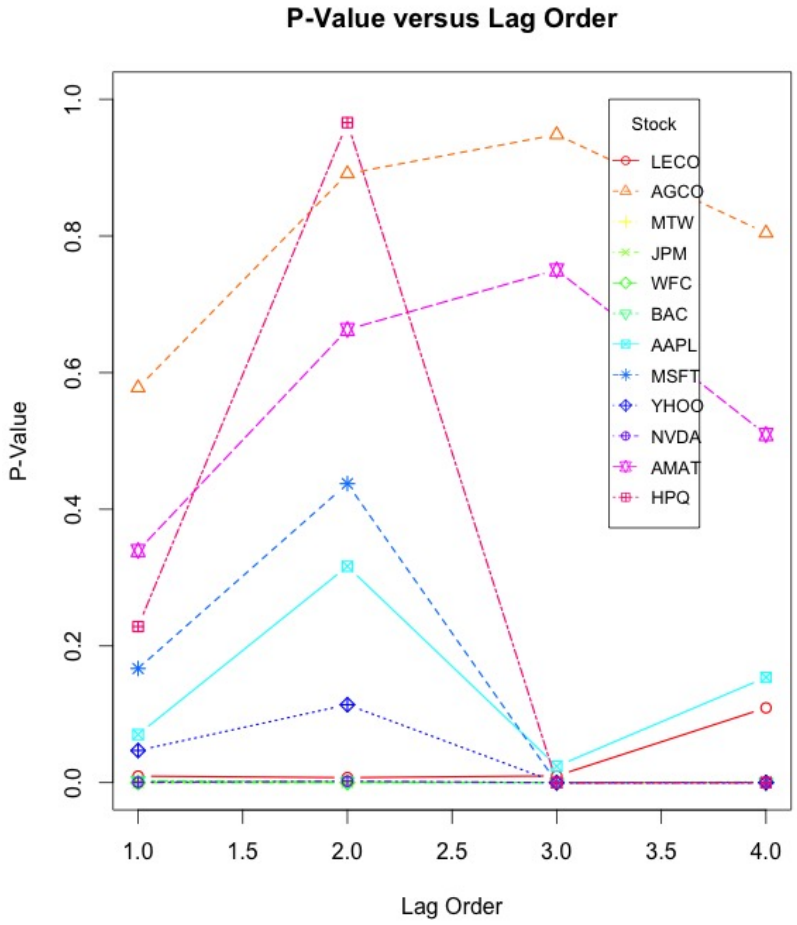

Figure 5.3: The plot contains the granger p-value results at various lags for the stocks in our portfolio. Notice how AGCO has p-values of greater than 0.5. This means that there is insigificant evidence to support the hypothesis that AGCO search volume can significantly forecase its volatility.

After running the Granger causality test for all stocks in our portfolio with Yang-Zhang volatility with $n = 3$ as the dependent variable and Google search volume the independent variable, both adjusted to be stationary, the following stocks have a Google search volume that could predict its Yang-Zhang volatility: MTW, JPM, WFC, BAC, AAPL, MSFT, YHOO,

NVDA, and HPQ.

The results support the view that for some of the stocks in our portfolio, the weekly Google search volume determined by its ticker symbol could have some element of predictability for the Yhang-Zhang volatility with $n = 3$. In the next section, a comparison will be made between several models to indicate how predicable Google search volume data can be.

5.3 Regression

After running the Granger causality analysis for all stocks and finding some combination of ticker symbol, volatility calculation, and order that is significant, we consider fitting a linear regression model to forecast volatility based on the search volume data. The stocks in table 5.3 lists stocks that are significantly caused by their Google Trends search volume ticker symbol.

Sector	Stock	Symbol
Technology	Apple Inc.	AAPL
	Microsoft Corporation	MSFT
	Yahoo! Inc.	YHOO
	NVIDIA Corporation	NVDA
	Hewlett-Packard Company	HPQ
Financial	JPMorgan Chase	JPM
	Wells Fargo & Co	WFC
	Bank of America Corp	BAC
Industrial	Manitowoc Company Inc	MTW

Table 5.3: A table of significant stocks that are Granger caused by their Google Trends search volume ticker symbol.

The data is split into training and testing sets. The training set consists of the first 80% of observations and the testing set consists of the last 20% of observations. Time series cross validation with a sliding window is used.

5.3.1 Null regression model

We fit the null regression model to the stocks in table 5.3. The model statistics are displayed in table 5.4.

Stock	MAE	MSE	RMSE	MAPE
MTW	0.2272	0.1006	0.3172	0.5061
JPM	0.1684	0.07451	0.273	0.5891
WFC	0.219	0.1454	0.3814	0.8196
BAC	0.2725	0.2256	0.475	0.9827
AAPL	0.1307	0.06571	0.2563	0.4377
MSFT	0.08328	0.01431	0.1196	0.3932
YHOO	0.1255	0.03979	0.1995	0.3682
NVDA	0.1415	0.04382	0.2093	0.2792
HPQ	0.09918	0.02077	0.1441	0.3883

Table 5.4: Null regression results in the training set for stocks.

If we look at the MAPE, we see that the null regression model did not really produce good results, as the values are around 40% and higher. The results of the null regression model for the testing dataset are in table 5.5.

Stock	MAE	MSE	RMSE	MAPE
MTW	0.1607	0.03418	0.1849	0.562
JPM	0.1113	0.01466	0.1211	0.7086
WFC	0.1628	0.02822	0.168	1.266
BAC	0.1554	0.02793	0.1671	0.8637
AAPL	0.3447	1.57	1.253	0.8339
MSFT	0.05505	0.004284	0.06545	0.3109
YHOO	0.0849	0.009872	0.09936	0.3697
NVDA	0.2038	0.04584	0.2141	1.031
HPQ	0.06867	0.0071	0.08426	0.3114

Table 5.5: Null regression results in the testing set for stocks.

Again, looking at the MAPE, we see that some models produced mean absolute percentage errors of over 100%. Although these models did not

behave particularly well, we now turn to see if models with the search volume and lags of volatility will perform better.

5.3.2 Simple linear regression

Below are results for the simple linear regression model with search volume, volatility, and search volume and volatility as predictors for the training and testing set.

Search volume model: Table 5.6 and table 5.7 list the results for the training and testing set for the search volume linear regression model. There is little to indicate that this model behaves better than the null model.

Stock	MAE	MSE	RMSE	MAPE
MTW	0.2335	0.09936	0.3152	0.5514
JPM	0.1807	0.07146	0.2673	0.6882
WFC	0.2376	0.1355	0.3681	0.9946
BAC	0.2915	0.2051	0.4528	1.155
AAPL	0.1258	0.06468	0.2543	0.4021
MSFT	0.08291	0.01363	0.1168	0.3999
YHOO	0.1225	0.03339	0.1827	0.3804
NVDA	0.1472	0.04055	0.2014	0.3228
HPQ	0.09751	0.0196	0.14	0.3855

Table 5.6: Simple linear regression results in the training set for stocks with search volume as predictors.

Stock	MAE	MSE	RMSE	MAPE
MTW	0.1845	0.04338	0.2083	0.6436
JPM	0.1388	0.02163	0.1471	0.8613
WFC	0.2017	0.04278	0.2068	1.529
BAC	0.1884	0.04289	0.2071	1.008
AAPL	0.3268	1.553	1.246	0.7363
MSFT	0.04958	0.003535	0.05946	0.2886
YHOO	0.0916	0.01139	0.1067	0.3999
NVDA	0.2526	0.06837	0.2615	1.256
HPQ	0.06627	0.006186	0.07865	0.3038

Table 5.7: Simple linear regression results in the testing set for stocks with search volume as predictors.

Volatility model: Table 5.8 and table 5.9 list the results for the training and testing set for the volatility linear regression model. There is little to indicate that this model behaves better than the null model.

Stock	MAE	MSE	RMSE	MAPE
MTW	0.2223	0.0884	0.2973	0.5301
JPM	0.1758	0.06839	0.2615	0.6669
WFC	0.2287	0.1246	0.353	0.9897
BAC	0.2811	0.1988	0.4458	1.138
AAPL	0.1175	0.04741	0.2177	0.3887
MSFT	0.07923	0.01277	0.113	0.3816
YHOO	0.119	0.0316	0.1778	0.371
NVDA	0.1383	0.03394	0.1842	0.3083
HPQ	0.0921	0.01735	0.1317	0.3698

Table 5.8: Simple linear regression results in the training set for stocks with search volume as predictors.

Stock	MAE	MSE	RMSE	MAPE
MTW	0.1756	0.03846	0.1961	0.5947
JPM	0.1393	0.02126	0.1458	0.8539
WFC	0.2031	0.04248	0.2061	1.538
BAC	0.1928	0.0402	0.2005	1.034
AAPL	0.3217	1.025	1.012	1.228
MSFT	0.04927	0.003294	0.0574	0.2762
YHOO	0.08482	0.009859	0.09929	0.3657
NVDA	0.2542	0.06737	0.2596	1.23
HPQ	0.05862	0.005078	0.07126	0.2699

Table 5.9: Simple linear regression results in the testing set for stocks with search volume as predictors

Search volume and volatility model: Table 5.10 and table 5.11 list the results for the training and testing set for the search volume and volatility linear regression model. There is little to indicate that this model behaves better than the null model.

Stock	MAE	MSE	RMSE	MAPE
MTW	0.2218	0.08832	0.2972	0.529
JPM	0.1765	0.06801	0.2608	0.6708
WFC	0.2317	0.1213	0.3483	1.003
BAC	0.2842	0.1914	0.4375	1.151
AAPL	0.1178	0.04674	0.2162	0.3912
MSFT	0.07917	0.01259	0.1122	0.3821
YHOO	0.1173	0.02985	0.1728	0.3685
NVDA	0.1381	0.0337	0.1836	0.3084
HPQ	0.0916	0.01715	0.131	0.3673

Table 5.10: Simple linear regression results in the training set for stocks with search volume and volatility as predictors.

Stock	MAE	MSE	RMSE	MAPE
MTW	0.1757	0.03849	0.1962	0.5951
JPM	0.1394	0.02133	0.146	0.8535
WFC	0.2022	0.04233	0.2057	1.521
BAC	0.1888	0.04058	0.2014	1.005
AAPL	0.3246	1.018	1.009	1.253
MSFT	0.04781	0.00313	0.05595	0.2684
YHOO	0.08677	0.009936	0.09968	0.3703
NVDA	0.2545	0.06749	0.2598	1.23
HPQ	0.0576	0.004868	0.06977	0.2669

Table 5.11: Simple linear regression results in the testing set for stocks with search volume and volatility as predictors.

5.3.3 Support vector regression

Search volume model: Table 5.12 and table 5.13 list the results for the training and testing set for the search volume support vector regression model. The MAPE values for these indicate that it behaves better than both the null model and linear regression model.

Stock	MAE	MSE	RMSE	MAPE
MTW	0.2228	0.1073	0.3275	0.444
JPM	0.1569	0.08015	0.2831	0.4399
WFC	0.2047	0.1548	0.3934	0.5763
BAC	0.2546	0.2369	0.4868	0.6798
AAPL	0.1175	0.06682	0.2585	0.3294
MSFT	0.07746	0.01485	0.1218	0.3181
YHOO	0.1153	0.03602	0.1898	0.3103
NVDA	0.1404	0.04323	0.2079	0.2759
HPQ	0.09205	0.02063	0.1436	0.3251

Table 5.12: Support vector regression results in the training set for stocks with search volume as predictors.

Stock	MAE	MSE	RMSE	MAPE
MTW	0.1314	0.02357	0.1535	0.4465
JPM	0.05917	0.005034	0.07095	0.397
WFC	0.089	0.009073	0.09525	0.7116
BAC	0.07202	0.006789	0.0824	0.4095
AAPL	0.2923	1.565	1.251	0.5441
MSFT	0.0472	0.003571	0.05975	0.2348
YHOO	0.06696	0.006422	0.08014	0.2806
NVDA	0.203	0.04537	0.213	1.024
HPQ	0.05881	0.005852	0.0765	0.2437

Table 5.13: Support vector regression results in the testing set for stocks with search volume as predictors.

Volatility model: Table 5.14 and table 5.15 list the results for the training and testing set for the volatility linear regression model. The MAPE values for these indicate that it behaves better than both the null model and linear regression model.

Stock	MAE	MSE	RMSE	MAPE
MTW	0.2136	0.09324	0.3054	0.4473
JPM	0.1547	0.07568	0.2751	0.435
WFC	0.2028	0.1378	0.3712	0.6143
BAC	0.2508	0.2164	0.4652	0.7198
AAPL	0.1118	0.0489	0.2211	0.3315
MSFT	0.07398	0.01378	0.1174	0.3072
YHOO	0.1117	0.0336	0.1833	0.3073
NVDA	0.133	0.03604	0.1899	0.2691
HPQ	0.08834	0.01821	0.1349	0.3215

Table 5.14: Simple linear regression results in the testing set for stocks with search volume and volatility as predictors.

Stock	MAE	MSE	RMSE	MAPE
MTW	0.1248	0.02064	0.1437	0.4199
JPM	0.05824	0.004853	0.06966	0.3843
WFC	0.09328	0.009769	0.09884	0.7366
BAC	0.0696	0.006613	0.08132	0.3982
AAPL	0.2913	1.034	1.017	1.067
MSFT	0.04122	0.002885	0.05371	0.2047
YHOO	0.05992	0.005134	0.07165	0.2488
NVDA	0.2088	0.04621	0.215	1.019
HPQ	0.05388	0.00465	0.06819	0.2293

Table 5.15: Support vector regression results in the training set for stocks with volatility as predictors.

Search volume and volatility model: Table 5.16 and table 5.17 list the results for the training and testing set for the search volume and volatility support vector regression model.

Stock	MAE	MSE	RMSE	MAPE
MTW	0.2125	0.09398	0.3066	0.4434
JPM	0.1546	0.07572	0.2752	0.4345
WFC	0.2024	0.1388	0.3726	0.6216
BAC	0.2507	0.2156	0.4643	0.7128
AAPL	0.1111	0.04853	0.2203	0.3325
MSFT	0.07394	0.01372	0.1171	0.3063
YHOO	0.1098	0.03197	0.1788	0.3048
NVDA	0.1327	0.03599	0.1897	0.269
HPQ	0.08778	0.01813	0.1346	0.3172

Table 5.16: Support vector regression results in the training set for stocks with search volume and volatility as predictors.

Stock	MAE	MSE	RMSE	MAPE
MTW	0.1237	0.02053	0.1433	0.4148
JPM	0.05826	0.004851	0.06965	0.3843
WFC	0.09248	0.009751	0.09875	0.7276
BAC	0.06887	0.006477	0.08048	0.3942
AAPL	0.293	1.037	1.018	1.097
MSFT	0.04139	0.002908	0.05393	0.2048
YHOO	0.06049	0.005141	0.0717	0.2516
NVDA	0.2081	0.04597	0.2144	1.012
HPQ	0.05304	0.004501	0.06709	0.2255

Table 5.17: Support vector regression results in the testing set for stocks with search volume and volatility as predictors.

5.4 Summary

The results indicate a few things. The first is that Google search volume data has some predictably element in them to forecast future volatility measures. This could be expected: whenever there is news about some company, more people are likely to search for the company. The second is that when using a support vector regression, the Google search volume data can provide an improvement to forecasting future volatility measures as compared to the null regression model. All in all, the results support the view that Google search volume data has some element of predicitiably in forecasting future volatility.

Chapter 6

Conclusion

6.1 Summary of findings

The results indicate that the search volume Granger causes volatility for some securities, or that some stocks search volume could predict volatility. When comparing the search volume model to the null model to forecast volatility, the MAPE showed that the search volume model performed better than the null model. Thus, the conclusion is that weekly Google search volume data has some predictably in forecasting stock volatility.

6.2 Comparison with related approaches

Thomas Dimpfl and Stephan Jank performed a similar analysis using search volume data and its effect on the DJIA historical volatility. Their results compared a naive model consisting of previous volatility values against a model consisting of previous volatility values and previous search volume values. The results show that the latter model with the search volume values increases mean squared error and other model statistics, but not by much. The results in this thesis reflect the results in their work.

6.3 Directions for future work

Other sources of data: Although Google search volume did not significantly forecast volatility, other researchers have used data from Twitter, forums, and the weather to significantly predict movements in the market. The process and research through this thesis will be useful when pursuing this route.

Volatility strategies: The naive model for forecasting volatility performed quite well. One could consider using options trading strategies to profit off

of volatility. Options such as a long straddle can be used to trade when an investor thinks highly volatile markets are likely, and short strangle can be used to trade when an investor thinks low volatile markets are likely. However, one must take into account risk management and transaction costs when designing trades.

Trading system: If one has a viable trading strategy, one could create a trading system that automatically connects to the market and execute trades. This area has a whole set of research questions to explore, such as optimal order frequency, transaction cost evaluation, and data warehousing and management, to name a few.

Bibliography

[1] A. Zeckman, "Google search engine market share nears 68%," November 2014. [Online]. Available: http://searchenginewatch.com/sew/study/2345837/google-search-engine-market-share-nears-68

[2] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *CoRR*, vol. abs/1010.3003, 2010. [Online]. Available: http://arxiv.org/abs/1010.3003

[3] A. Chyan, T. Hsieh, and C. Lengerich, "A stock-purchasing agent from sentiment analysis of twitter," Stanford University, Tech. Rep., 2011. [Online]. Available: http://cs229.stanford.edu/proj2011/ChyanHsiehLengerich-A_Stock-Purchasing_Agent_from_Sentiment_Analysis_of_Twitter.pdf

[4] R. Chakoumakos, S. Trusheim, and V. Yendluri, "Automated market sentiment analysis of twitter for options trading," Stanford University, Tech. Rep., 2011. [Online]. Available: http://cs229.stanford.edu/proj2011/TrusheimChakoumakosYendluri-Automated_Market_Sentiment_analysis_of_Twitter_for_Options_Trading.pdf

[5] V. Kuleshov, "Can twitter predict the stock market?" Stanford University, Tech. Rep., 2011. [Online]. Available: http://cs229.stanford.edu/proj2011/Kuleshov-CanTwitterPredictTheStockMarket.pdf

[6] D. Debbini, P. Estin, and M. Goutagny, "Modelling the stock market using twitter sentiment analysis," Stanford University, Tech. Rep., 2011. [Online]. Available: http://cs229.stanford.edu/proj2011/DebbiniEstinGoutagny-ModelingTheStockMarketUsingTwitterSentimentAnalysis.pdf

[7] E. Hsu, S. Shiu, and D. Torczynski, "Predicting dow jones movement with twitter," Stanford University, Tech. Rep., 2011. [Online]. Available: http://cs229.stanford.edu/proj2011/HsuShiuTorczynski-PredictingDowJonesMovementWithTwitter.pdf

[8] A. Mittal and A. Goel, "Stock prediction using twitter sentiment analysis," Stanford University, Tech. Rep., 2011. [On-

line]. Available: http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf

[9] R. Chen and M. Lazer, "Sentiment analysis of twitter feeds for the prediction of stock market movement," Stanford University, Tech. Rep., 2011. [Online]. Available: http://cs229.stanford.edu/proj2011/ChenLazer-SentimentAnalysisOfTwitterFeedsForThePredictionOfStockMarketMovement.pdf

[10] G. Gidfalvi, "Using news articles to predict stock price movements," University of California, San Diego, Tech. Rep., 2001. [Online]. Available: http://cseweb.ucsd.edu/~elkan/254spring01/gidofalvirep.pdf

[11] K. Lee and R. Timmons, "Predicting the stock market with news articles," Stanford University, Tech. Rep., 2007. [Online]. Available: http://nlp.stanford.edu/courses/cs224n/2007/fp/timmonsr-kylee84.pdf

[12] R. P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news: The azfin text system," *ACM Trans. Inf. Syst.*, vol. 27, no. 2, pp. 12:1–12:19, Mar. 2009. [Online]. Available: http://doi.acm.org/10.1145/1462198.1462204

[13] T. Preis, H. S. Moat, and H. E. Stanley, "Quantifying trading behavior in financial markets using google trends," *Sci. Rep.*

Academic Vita

Christopher Siergiej

200 W College Ave Apt 4

State College, PA 16801

cvs5202@psu.edu

Education

Pennsylvania State University, Schreyer Honors College, 2015

M.A.S. Statistics

B.S. Computer Science

B.S. Mathematics

Employment

Cisco Systems, Software Development Engineer Intern, 2014

Deloitte Consulting, Human Capital Actuarial Analyst Summer Scholar, 2013

IBM, Software Development Engineer in Test Intern, 2012

Activities

Penn State Blue Band, Drum Major

Awards

John K. Tsui Scholarship

Gregory and Robert Stock Endowed Scholarship

Certifications

SOA Exam 1/Probability

SOA Exam 2/Financial Mathematics