April 2016

Modeling Investor Sentiment to Protect Against Downside Market Risk

Dylan Wadsworth McCarthy
Worcester Polytechnic Institute

Modeling Investor Sentiment to Protect Against Downside Market Risk

A Major Qualifying Project
Submitted to the faculty of
WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the
Degree of Bachelor of Science
by

Dylan Wadsworth McCarthy
dwmccarthy@wpi.edu

Date: April 27, 2016

Approved:

Professor Dimitrios Koutmos, Advisor

Abstract

The goal for this Major Qualifying Project was to research, examine, create, and modify existing algorithmic trading strategies in order to account for investor sentiment. We were able to quantify investor sentiment through the use of publicly available pageview data from prominent Internet organizations. With this data, we tested several strategies, some of which were market-blind in the sense that they traded only on the basis of sentiment indicators. While we do not refute that sentiment indicators are likely (in some respect) driven by market indicators such as asset pricing and volatility, finding a successful strategy that trades solely on the basis of sentiment indicators is a challenge.

One of the most successful strategies we found, and the strategy we will be focusing on, is able to achieve returns close to that of a buy-and-hold strategy during stable and bull market periods, and does a phenomenal job of mitigating downside market risk in bear markets. The potential impacts of utilizing such a strategy in the real world are realizable: while many investors are experiencing tremendous portfolio depreciation, an investor utilizing a sentiment-based trading strategy could potentially insulate their portfolio from bear markets, suffering minimal losses.

Acknowledgments

I would like to thank Professor Dimitrios Koutmos of Worcester Polytechnic Institute for advising this project, tutoring me in market and wealth-management strategies, and for sparking my interest in economic and financial analysis. It is due to his commitment and persistence that I have found a desire to pursue a career in investment, international markets, or wealth management.

Additionally, I would like to thank Wikipedia for fully publishing pageview data and other metrics, on which we relied to quantify sentiment.

Finally, I would like to thank Worcester Polytechnic Institute's Robert A. Foisie School of Business for a comprehensive and immersive education and for the ability to work on this project.

Contents

List of Figures

List of Tables

1 Introduction

1.1 What is the Major Qualifying Project?

The Major Qualifying Project (MQP) is a mandatory capstone project to be completed by all undergraduate students in their final year of study at Worcester Polytechnic Institute (WPI). This project is intended to focus on a specific interest in the student's field of study and lead to a learning experience that is both educational, interesting, and potentially useful in the student's future career. It is preceded by an Interactive Qualifying Project, an introductory version of the Major Qualifying Project that is completed in the student's second-to-last year of study.

1.2 Motivations and Problems it Seeks to Solve

We chose to focus our Major Qualifying Project on identifying alternative ways in which the wealth of information available electronically can be used to assist those who trade in both national and international stock markets. As the information age came to fruition throughout the 1990's and early 2000's, markets became less segmented and investors began to realize that regardless of their location, they could communicate and trade with companies and other investors around the world. As investing became further globalized and integrated with technology, traders began to further scrutinize potential investments as those investments may be in another state or even country with rules and regulations different than their own.

Two prevailing equity analysis techniques are prevalent in the equity investing world: fundamental and technical analysis. Fundamental analysis focuses on the company itself: for example, fundamental analysis of General Electric's stock (`NYSE:GE`) would focus on how the company is

run, the management structure, and would include detailed analysis of company financials. Warren Buffet is a famous billionaire investor that generally follows the fundamental investing strategy. Ideally, a complete fundamental analysis would give a potential investor an idea of whether the company is currently overvalued or undervalued based on the evaluating factors.

Technical analysis focuses on finding and exploiting pricing differences in the market. While fundamental analysis will look at the company as a whole, technical analysis will evaluate the company on measures of price, volume, and other market indicators. A technical analysis approach will compare these indicators to historical data in an attempt to gain insight into where the price of the security is headed. George Soros is a well known technical investor.

Although these analysis tools have been in use for many decades, I was curious about analyzing equities based not on projected future performance or financial and market viability, but based on how investors perceive these stocks. A company without a positive future outlook can still have excellent stock price performance if investors believe the company holds future value.

Indeed, pump and dump schemes were very common in unregulated markets with penny stocks where investors would be tricked into believing the company has great potential and is greatly undervalued, and would buy shares, inflating the share price, until the perpetrators sold their shares, causing a collapse in the share price. Jonathan Lebed became famous for perpetrating many pump and dump schemes while he was still in high school.[1] These schemes have nothing to do with the underlying security and everything to do with tricking the investors into thinking they have a chance to invest in a severely undervalued equity.

I reasoned that since investors ultimately exhibit control over the share price through the buy-

[1] http://www.nytimes.com/2001/02/25/magazine/25STOCK-TRADER.html?pagewanted=all

ing and selling of shares, it would be fascinating to be able to quantify the investor's sentiment about the securities they invest in.

For future use in this paper, the distinction must be made between informed investors and uninformed investors. From this point on, the term 'informed investors' will refer to those those who have superior knowledge, access to specialized investing and equity analysis resources, and the experience to use them. Examples of informed investors include multi-billion dollar firms such as Goldman Sachs and UBS, as well as individuals ranging from Warren Buffet and Carl Icahn to an experienced trader with a proven track record. Uninformed investors, meanwhile, are generally categorized by those with less than 10 years of investing experience, lack of access to specialized tools, or an undeveloped or untested investing mentality.

It will become apparent that the measures of investor sentiment that we utilize mainly classify the sentiment of uninformed investors, as informed investors generally use specialized tools for equity analysis rather than traditional web and encyclopedia searches.

Additionally, the efficient market hypothesis (EMH) states "it is impossible to 'beat the market' because stock market efficiency causes existing share prices to always incorporate and reflect all relevant information".[2] Some believe that the efficient market hypothesis is always true, while others disbelieve it entirely, and still others believe that at times the market is efficient and at times it is inefficient. Studying investor sentiment will allow us to understand whether the EMH holds true throughout different market periods, because if all relevant information is incorporated in the share price, we could not leverage greater returns.

[2] http://www.investopedia.com/terms/e/efficientmarkethypothesis.asp

1.3 Global and Multidisciplinary Importance

The global importance of this Major Qualifying Project lies in the fact that hundreds of millions of people worldwide look to investments in order to leverage their financial situation, provide income, or otherwise financially benefit themselves. Traditional investment strategies call on diversifying investments in order to lower potential downside risk and insulate against volatility, regardless of the investment medium. However, recent research provides data indicating that overall correlation between markets tends to greatly increase during depressed economies: in a 2000 journal article, Figlewski et al. study whether the leverage effect applies throughout bull and bear markets.[3] Ultimately, this exacerbates the need for novel investment strategies that can protect investors against downside market risk.

This Major Qualifying Project was, without question, multidisciplinary in nature. Designing and testing algorithmic trading strategies requires knowledge of computer science in order to gather data and implement strategies. Studying human sentiment requires a psychological approach. Finally, understanding equity trading and analysis requires financial and investment management knowledge.

Search volume is utilized in a multidisciplinary manner as well: With publicly available search volume, people took interest in modeling the data in order to predict certain outcomes. In 2009, employees of Google and the Centers for Disease Control and Protection (Ginsberg et al.) published a paper outlining their attempt to use search volume data in order to detect influenza epidemics.[4] In an October 2010 edition of the National Academy of Sciences journal (PNAS), an

[3] http://papers.ssrn.com/sol3/papers.cfm?abstract_id=256109

[4] http://static.googleusercontent.com/media/research.google.com/en//archive/papers/detecting-influenza-epidemics.pdf

article by Goel et al. showed that search volume can be successfully used to "predict their [consumers] collective future behavior days or even weeks in advance".[5]

[5] http://www.pnas.org/content/107/41/17486

2 Recent Interest in Information Demand

The global financial markets are a huge draw for many individuals, whether they are investing as a hobby or as a profession. Recent startup companies have lowered the barriers to entry on the equity investment markets. While companies such as Robinhood[6] offer commission-free trading, other companies like Acorns[7] automatically invests an individual's spare change and then manages the investments across different sectors to achieve the desired returns. In a typical day on Wall Street, some portfolios will gain vale and some will lose value, but there is no question that a great many people and firms participate in both national and international equity trading.

Whether the investor is informed or uninformed, it would be silly and downright careless to invest without prior research. The early 2000's ushered in the Information Technology age, where information became readily available to anyone with access to a computer. Investors that once used the telephone for research began to migrate to the Internet as it was faster and easier to use.

With the advent of the Internet came tools and resources dedicated to equity trading. The Bloomberg Terminal[8], for example, was invented in 1982 and allowed investors to access the latest market data in seconds, something that was previously impossible. It is important to note that resources such as the Bloomberg Terminal were generally only accessible by informed investors (backed by large firms) or those with substantial resources; indeed, the Bloomberg Terminal can cost upwards of $20,000 USD per year.

As services such as Google and Wikipedia became available, uninformed traders turned to these tools in order to research potential investments. As the total number of Internet users began

[6] https://www.robinhood.com
[7] https://www.acorns.com
[8] http://www.bloomberg.com/professional/hardware

to grow, these services began to accumulate detailed records on what people were searching for. In 2006, Google launched Google Trends[9], which allows for users see certain popular search terms, as well as lookup a search term in order to see the term's historical search volume. During its inception in 2001, Wikipedia, as a part of their transparency initiative chose to publish much of their dataset[10], which includes article traffic counts.

With such an extensive set of data available, many have proposed modeling search volume in order to gain a competitive edge on the global stock markets.

2.1 Literature Review

The reviewed literature falls into five distinct categories based on the strategies and data used.

2.1.1 Returns

Quantifying Trading Behavior in Financial Markets using Google Trends: *Tobias Preis, Helen Susannah Moat, and H. Eugene Stanley*

This journal article, published in an April 2013 article of Nature's Scientific Reports, uses Google Trends data to examine 98 terms that are in some way related to the financial markets, such as 'debt'. Their strategy buys or sells (shorts) the Dow Jones Industrial Average (DJIA) depending on the difference between the Google Trends search volume data with historical data.

Their results were very positive, with the Google Trends strategy on the keyword 'debt' returning 326% across an 8 year span versus a 16% buy-and-hold return. Furthermore, the authors detailed the individual terms and their cumulative returns in order to compare each term to the

[9] `https://www.google.com/trends`
[10] `https://en.wikipedia.org/wiki/Wikipedia:Statistics`

others. The authors concluded that utilizing large-scale data such as sentiment from search volume could lead to exciting future possibilities.

To note, the Google Trends data is now accessed via a different interface, and Google has taken steps to normalize the data and has made it inconvenient to download or otherwise parse the data for programmatic analysis, which made Google Trends as a data source a less viable option for the focus of our strategy.

Quantifying the semantics of search behavior before stock market moves: *Chester Curme, Tobias Preis, H. Eugene Stanley, and Helen Susannah Moat*

Published in late 2013, this article has shared authors with the first article, Quantifying Trading Behavior in Financial Markets using Google Trends. This strategy also used Google Trends in order to gauge sentiment through search volume, but the terms to be examined in Google Trends were generated via a latent Dirichlet allocation model, which was set to determine 100 different topics with 30 words within each topic that best represent that topic, sourced from Wikipedia. The topics were labelled through the use of Amazon Mechanical Turk, a quasi human-outsourcing project.

For those topics which had enough search volume (so that Google Trends would display them), the trading strategy would then buy or sell (short) the S&P 500 Total Return index (SPXT) based on historical difference of search volume for the terms. The authors concluded that there is no "robust" connection between a diverse set of topics and stock market moves.

To note, the authors plotted the Google Trends terms versus the term 'google' in order to normalize the data between terms, as the Google Trends engine shows resulting volume on a

scale from 1-100 for individual terms that are chosen.

In Search of Attention: *Zhi Da, Joseph Engelberg, and Pengjie Gao*

This article was published in 2011 in the Journal of Finance by the American Finance Association. These authors also use Google Trends as a measure of search volume, but in this case their strategy uses the ticker symbol as the measure of search volume rather than a proxy measure of volume such as the prior articles used (e.g. Wikipedia). This strategy measures the search volume from each ticker in the Russel 3000 exchange traded fund, an ETF that is generally regarded as a market benchmark.[11] The strategy also includes measures to compensate for noise and potential interference in the data by filtering out data and results that match certain criteria.

For this search volume research, the authors found that there was a positive correlation between search volume and short-term price increases. Ultimately, they concluded that higher search volume results in higher asset prices for the following two weeks, with prices reversing within 1 year.

The article goes on to explore and analyze the relationship between search volume and companies that are going to have an initial public offering (IPO) in the near future. Other topics the authors research include classifying those investors that are reflected in the search volume measures, and other applications for search data.

2.1.2 Risk

Information demand and stock market volatility: *Nikolaos Vlastakis and Raphael N. Markellos*

This article, published in the Journal of Banking & Finance in 2012, focuses on large-cap com-

[11] `http://www.investopedia.com/terms/r/russell_3000.asp`

panies on the NYSE and NASDAQ. Similar to the previous articles, the authors gauge information demand through Google Trends search volume data (using S&P 500 as the Google Trends keyword).

Google Trends search volume is obtained for the 30 largest stocks on the NYSE and NASDAQ. Rather than predicting future returns or market moves like the previous research, this article concludes that information demand and historical volatility, volume, and implied volatility, are closely and positively linked. The authors also find that the relationships previously stated are strengthened during bull markets.

2.1.3 Niche Finance

Google Internet search activity and volatility prediction in the market for foreign currency: *Geoffrey Peter Smith*

Published in the Elsevier Finance Research Letters journal in early 2012, Smith also studies Google Trends data in order to determine search volume for certain keywords and correlate search volume to future FOREX volatility. Initially, the author researched search volume for specific searches that included 'economic' and 'crisis', or 'financial' and 'crisis'.

Smiths' results show that certain combinations of keywords outperform the generalized autoregressive conditional heteroscedasticity (GARCH(1,1)) model. Ultimately, the author shows that search volume could be used as a model for predicting future volatility in FOREX markets.

Google attention and target price run ups: *Antonios Siganos*

This article examines certain price changes in companies before a merger is announced to the public. Published in late 2012 in the International Review of Financial Analysis, Siganos attempts

to determine whether or not search volume can predict future mergers, and if search volume can explain price run ups.

The author concludes that search volume derived from Google Trends tends to signal a merger due to a surge in search volume. This largely occurs before other sources report the merger publicly. The article goes on to say that this increase in search volume cannot explain the price run ups observed. The author notes that this information should be interesting to regulatory agencies as there are regulations that attempt to control target price run ups.

2.1.4 Sentiment

The Sum of All FEARS Investor Sentiment and Asset Prices: *Zhi Da, Joseph Engelberg, and Pengjie Gao*

This article is longer than those previously reviewed, and attempts to build an index from Google Trends search volume data. It was published in late 2014 in The Review of Financial Studies. The index is constructed with search volume of keywords that exemplify investor fears, including terms such as recession and bankruptcy. The index is called FEARS (Financial and Economic Attitudes Revealed by Search) and is meant to aggregate investor sentiment toward the aforementioned keywords.

The authors conclude that their FEARS index can predict 'aggregate market returns', price reversals in the short term, volatility changes, and flows of funds from equity funds into bond funds.

Forecasting Abnormal Stock Returns and Trading Volume Using Investor Sentiment: Evidence from Online Search: *Kissan Joseph, M. Babajide Wintoki, and Zelin Zhang*

This article, published in late 2011 in the International Journal of Forecasting, uses Google Trends to examine search volume for certain stock tickers in an attempt to predict 'abnormal' returns and volume. The authors attempt to prove that stock ticker search volume can reliably be used as a measure of investor sentiment.

The findings show that search volume for stock tickers can indeed predict abnormal returns for stocks on the S&P 500 as well as inflated volume. The authors state that their results are similar to those found by Da et al, a journal article also reviewed earlier in this document. The prior article focused on Russel 3000 firms while this article focuses on S&P 500 firms, with similar results.

2.1.5 Wikipedia

Quantifying Wikipedia Usage Patterns Before Stock Market Moves: *Helen Susannah Moat, Chester Curme, Adam Avakian, Dror Y. Kenett, H. Eugene Stanley, and Tobias Preis*

This article is the second reviewed article that sources data from Wikipedia, as most of the reviewed literature sources data from Google Trends. Published in Nature's Scientific Reports journal in mid 2013, the authors design a strategy for trading the DJIA based on Wikipedia page view counts and page edit counts from 1) DJIA companies, 2) actors & filmmakers, and 3) financial topics.

Their results show that Wikipedia page views for DJIA companies and financial topics tend to increase before pricing falls. The actors and filmmakers page views show no noticeable correlation with market moves. Furthermore, Wikipedia page edits don't show noticeable predictive power in any of the test cases.

2.2 Gaps in Existing Studies

We performed extensive literature review in an effort to fully understand the previous studies in investor sentiment and information demand. It was evident to us that most of the studies sourced data from Google Trends, and for good reason. Google is the most popular website in the world[12], and for good reason, allowing anyone to search virtually anything with instant results. With the absolutely massive user base, as of 2012 Google was processing more than 1.2 trillion searches per year[13], or an average of 75 searches for each person on the planet Earth.

However, there are some downfalls to the Google Trends strategy. Although Google publicizes some search data, is is usually normalized and publicized as a percentage rather as a discrete number. Searching the keyword 'WPI' on Google Trends will show historical data, but the date with the highest search volume for 'WPI' will appear as 100%, while the other dates will be scaled from that measurement. Similarly, it becomes challenging to compare multiple keywords, as the values for each keyword will be scaled off the keyword with the highest search volume. Some studies have opted to always include 'Google' in the keywords, so they can more accurately compare multiple keywords.

Existing studies mainly focus on the mathematical analysis behind potential data sources and strategies, but many fail to test if such a strategy would be viable in the real-world financial markets that exist outside the theoretical.

We decided to base the focus of our study on the final article (Quantifying Wikipedia Usage Patterns Before Stock Market Moves, Moat et al.) in an effort to further research Wikipedia's data-

[12] `http://www.alexa.com/topsites`
[13] `http://www.internetlivestats.com/google-search-statistics`

set to see if it too could provide similar results to the results other journal articles found with the Google Trends data-set. Our study will incorporate real-world analysis where we will test if such a strategy is at all viable in financial markets.

2.3 Existing Gaps in Financial Decision-making

Current financial analysis techniques mainly focus on analysis of the equity and its underlying company. A decision is reached as to what trade should be made (or if a trade should be made at all), and then the trade is executed. Furthermore, we cannot know what trade an investor will make until the trade is completed. Although these techniques have worked well for the past 100 years, as technology continues to change and evolve, it would seem fitting to incrementally innovate on the traditional analysis techniques.

Computers have enabled trades to take place at the speed of light, and billions of dollars can change hands before the human mind can process what is going on. Although the financial markets are being automated and digitized, human decision making still greatly affects what occurs in the day-to-day operation of the markets. As data that enables investor sentiment analysis becomes more readily available and in larger quantities, it would seem to be short-sighted to continue to focus on traditional equity analysis without evolving with the times.

Quantifying investor sentiment will allow us to not only understand and potentially predict what goes on in an investor's mind before the trade, but would also allow existing economic models to be refined now that there is data to represent the human element.

3 Sample Data

3.1 Data Sources

All raw data is sourced from freely available sources, and we attempted to source it directly from the publisher, or from a reputable company, which would enable us to rely on the integrity of the data.

3.1.1 Investor Sentiment

We chose to quantify investor sentiment through Wikipedia page view data. It is important to note that both Google Trends and Wikipedia do not provide detailed search volume data - in fact, Google Trends only provides a percentage value while Wikipedia provides an integer value that represents page views. As these companies do not publish detailed data, we must use this generalized data with the understanding that there may be significant levels of error in the data: obviously, not every page view is that from an investor or trader, and may be from a student performing research, for example.

Important Note: The Wikipedia page view data set is not fully complete. Some month sets are missing several days while some sets are missing nearly an entire month. This posed many problems for our tests as we needed complete data in order to test strategies over a length of time. For this reason, we chose the period from August 2008 (08/2008) to December 2014 (12/2014)

The Wikipedia data is sourced from the Wikipedia data-set, which contains page view data:

`https://dumps.wikimedia.org/other/pagecounts-raw/`

Rather than downloading large data dumps, the page view data is published online by Domas

Mituzas, a prior Wikipedia developer. It is available via a website or programmatically through a JSON call:

```
http://stats.grok.se
```

The web interface is presented in Figure 1 below.

Figure 1: Wikipedia page views interface

The data returned from a JSON call follows the following format:

```
{"daily_views": {"2013-08-26": 8062, "2013-08-27": 7944,
   "2013-08-28": 6844, "2013-08-19": 5587, "2013-08-18": 4799,
   "2013-08-31": 4872, "2013-08-29": 7028, "2013-08-15": 5820,
   "2013-08-14": 6476, "2013-08-17": 5432, "2013-08-16": 6315,
   "2013-08-11": 5676, "2013-08-10": 5562, "2013-08-13": 6198,
   "2013-08-12": 6419, "2013-08-30": 6254, "2013-08-20": 6248,
```

```
"2013-08-21": 7328, "2013-08-22": 7712, "2013-08-23": 12194,
"2013-08-24": 8316, "2013-08-25": 6721, "2013-08-08": 5874,
"2013-08-09": 5582, "2013-08-06": 5866, "2013-08-07": 6006,
"2013-08-04": 4819, "2013-08-05": 5371, "2013-08-02": 5912,
"2013-08-03": 5274, "2013-08-01": 6215}, "project": "en", "month":
 "201308", "rank": 857, "title": "Microsoft"}
```


3.1.2 Market Data

All market data was sourced from Yahoo! Finance[14], a reputable source for financial market data that has been in existence for many years. Historical data can be downloaded in CSV (comma separated values) format that includues historical data ranging from the first day the ticker began trading up until the most recent trading day. The data includes the trading date, the volume, and the open, high, low, close, and adjusted close prices. The CSV can be downloaded from the bottom of the Historical Prices page of nearly any ticker:

`http://finance.yahoo.com/q/hp?s=AAPL+Historical+Prices`

Data from the CSV file will be structured as follows in Table 1.

[14] `http://finance.yahoo.com`

Date	Open	High	Low	Close	Volume	Adj. Close
2016-04-22	105.010002	106.480003	104.620003	105.68	33477100	105.68
2016-04-21	106.93	106.93	105.519997	105.970001	31356400	105.970001
2016-04-20	106.639999	108.089996	106.059998	107.129997	28666900	107.129997
2016-04-19	107.879997	108.00	106.230003	106.910004	32292300	106.910004
...	...	...	...	...	...	...
n	n_{open}	n_{high}	n_{low}	n_{close}	n_{volume}	$n_{adjclose}$

Table 1: Sample Yahoo Finance CSV Data

3.2 Software Framework

Throughout this project, we utilized Ruby as our software analysis framework. Ruby is extremely expandable through the use of "gems", which are generally open-source and can add functionality in many areas, from graphing to advanced computation.

Ruby was run in a virtualized environment on a Windows host, running an Ubuntu Desktop 15.10 64-bit instance, which was allocated 4GB of memory, 30GB of storage space, and on-demand use of the local processor.

The Wikipedia page view data and the Yahoo Finance ticker data were programmatically downloaded and saved in a local database, which was connected to the Ruby instance. This enabled us to query the data locally which significantly sped up the processing time.

4 Analytical Framework

4.1 Modeling Technique

We designed several different strategies, but they all had the same underlying framework. All strategies were quasi-dynamic in nature, meaning the strategy constraints are hard-coded but the decisions are made based on changing input variables.

Building on research from Moat et al., we decided to design a similar base strategy. At the beginning of each week, page view data from the past *t+1* weeks is analyzed. Weekly averages are computed from the daily Wikipedia data and are subsequently compared. Weekly averages for weeks 1 through *t* are grouped and then averaged to provide a historical weekly average page view count. This is compared to the weekly average of week *t+1*, which is defined as the previous week average.

Should the historical weekly average be greater than the previous week average, the strategy will instruct a BUY-SELL transaction which will buy the Dow Jones Industrial Average (DJIA) on the first market day of the current week, and will sell the DJIA to close the transaction on the first market day of the week *n* weeks from now.

Should the historical weekly average be less than the previous week average, the strategy will instruct a SELL-BUY transaction which will sell (short) the DJIA on the first market day of the current week, and will buy the DJIA to close the transaction on the first market day of the week *n* weeks from now.

In a generalized form:

If t = *number of past weeks*, n = *weeks to hold transaction open*

$$\frac{\sum_1^t Wk_i}{t} - Wk_0 < 0 \implies \text{sell (short) at } Wk_0 \text{ then buy at } Wk_n \tag{1}$$

$$\frac{\sum_1^t Wk_i}{t} - Wk_0 > 0 \implies \text{buy at } Wk_0 \text{ then sell at } Wk_n \tag{2}$$

One might ask: why choose these constraints? Based on the reviewed literature, surges in information demand (search volume) generally cause nearly immediate price increases followed by short-term price reversals. These equations generalize the assumption that if more people are searching for (or looking up) a keyword or article, the corresponding equity will rise in price and then have a short-term price reversal. Essentially, we are betting against the surge in price caused by heightened information demand. As past research has shown, the price tends to reverse quickly. By shorting the stock when people have been purchasing it and buying it back a certain number of weeks later, we are essentially counting on the fact that the price will reverse. We also buy a stock when the information demand falls, as we believe the price will rise when information demand begins to rise.

It is important to note that this quasi-dynamic strategy is unlikely to be as successful as a fully dynamic strategy. A fully dynamic strategy might, for example, buy the stock and continue to hold it until information demand begins to surge, and when prices rise as a result of the surge in information demand, sell the stock. Future research should focus on building fully dynamic strategies, as they can react directly to market conditions, without having hard-coded variables

such as our t weeks and n weeks to hold the transaction open.

As a general rule, unless otherwise specified, we used $t = 3$ and $n = 1$ based on calculations in Moat et al.

Throughout our strategies, we report both arithmetic and geometric returns. Many money management firms tend to use arithmetic mean returns, while many academics promote the use of geometric mean returns in financial documents. Generally, the arithmetic mean tends to overstate the value while geometric mean tends to understate the value of portfolios.[15][16]

4.2 Trading Strategies

Over the course of this project, we built and tested several trading strategies as we explored the wealth of information that is provided through market data as well as information demand data. We will highlight the different set of strategies we tried, and go into more detail about the most successful strategy, which is also the strategy we spent the most time designing.

4.2.1 Buy and Hold

The first strategy we built modeled a simple buy and hold strategy where the DJIA is bought on the beginning date and sold on the ending date. This strategy was very straightforward and was constructed in order to compare our sentiment-based trading strategies to how the market would perform over a certain time period.

[15] http://www.investopedia.com/ask/answers/06/geometricmean.asp
[16] http://www.meketagroup.com/documents/Modified%20Geometric%20Mean%20WP.pdf

4.2.2 Wikipedia & Dow 30

For our sentiment-based trading, we started with a basic strategy of trading the DJIA based on measures of investor sentiment from each of the Dow 30 companies. The Wikipedia page views for each of the Dow 30 companies were collected and averaged together, and then historically averaged, as explained in the Modeling Technique section. We reasoned that this was a good starting point as it dealt with each of the companies directly and involved correlating company page view data with market movements.

After we determined the overall weekly average for the total 26 companies, we followed the general framework set out in the Modeling Technique section. This framework involves trading the DJIA on weekly intervals based on increases or decreases in page view counts on the corresponding 26 Wikipedia articles.

It should be noted that the Dow Jones Industrial Average (DJIA) is comprised of 30 different large-cap companies traded on the stock exchange. The Wikipedia pages we chose for the companies that comprise the DJIA numbers 26 companies as we chose companies that remained in the DJIA through the period of our study, and excluded companies that either dropped out or entered the DJIA.

The 26 companies that remained in the DJIA from 08/2008 to 12/2014 are listed in Table 2.

This strategy did not successfully generate meaningful or substantive returns on the market. We realized that it's likely there was a lot of error from the Wikipedia data itself, as many people would visit Microsoft's Wikipedia page, for example, to look at their latest products rather than investigating the company as a potential target for investment. Our returns for this strategy are

No.	Company	Ticker
1	3M	NYSE: MMM
2	American Express	NYSE: AXP
3	AT&T	NYSE: T
4	Boeing	NYSE: BA
5	Caterpillar Inc.	NYSE: CAT
6	Chevron Corporation	NYSE: CVX
7	Du Pont De Nemours And Co	NYSE: DD
8	Exxon Mobil Corporation	NYSE: XOM
9	General Electric	NYSE: GE
10	Goldman Sachs Group Inc	NYSE: GS
11	International Business Machines Corp.	NYSE: IBM
12	Intel Corporation	NASDAQ: INTC
13	Johnson & Johnson	NYSE: JNJ
14	JPMorgan Chase & Co.	NYSE: JPM
15	McDonald's Corporation	NYSE: MCD
16	Merck & Co., Inc.	NYSE: MRK
17	Microsoft Corporation	NASDAQ: MSFT
18	Pfizer Inc.	NYSE: PFE
19	Procter & Gamble Co	NYSE: PG
20	The Coca-Cola Co	NYSE: KO
21	Home Depot Inc	NYSE: HD
22	Travelers Companies Inc	NYSE: TRV
23	Walt Disney Co	NYSE: DIS
24	United Technologies Corporation	NYSE: UTX
25	Verizon Communications Inc.	NYSE: VZ
26	Wal-Mart Stores, Inc.	NYSE: WMT

Table 2: Dow Jones Industrial Average Components from 08/08 to 12/14

summarized in Table 3. Note that just removing the highest company by Wikipedia page view volume from the strategy reversed the returns from negative to positive. This gave us more certainty that the popular companies were causing our strategy to trade with error, and at this point we decided to pivot to another strategy.

Strategy	Trading Dates	Arithmetic Return	Geometric Return
Base			
	08/2008 - 12/2014	-0.5868	-0.7084
	08/2008 - 11/2009	-0.4108	-0.4763
	08/2008 - 03/2009	-0.1500	-0.2060
Base w/o MSFT			
	08/2008 - 03/2009	0.2870	0.2306

Table 3: Wikipedia & Dow 30 Strategy Returns

4.2.3 Wikipedia & Economic Terms

Our final strategy involved the same framework outlined in the Modeling Technique section, but trading based on the page view count of general economic terms. In some of the literature review, it was suggested that general economic terms could be used to approximate the market. First, we chose several negative economic terms that would likely be more popular during bear markets: fear, recession, and unemployment.

The idea was to trade the DJIA using the page views for these pages as a trading signal. The original trading strategy can still be adopted, as we can assume that less interest in these pages can equate to bull or stable markets. If these pages experience a spike or stable increase in popularity, we can assume that these terms have some basis in the current economic and financial climate.

The returns are summarized below in Table 4. Fear was one of the most volatile results - performing extraordinarily in bear markets and poorly in bull markets. This is, to some extent, to be expected as we designed the strategy to perform during down markets.

Strategy	Trading Dates	Arithmetic Return	Geometric Return
fear			
	08/2008 - 12/2014	0.4243	0.3014
	08/2008 - 11/2009	0.7426	0.6754
	08/2008 - 03/2009	0.8688	0.8112
recession			
	08/2008 - 12/2014	0.6338	0.5108
	08/2008 - 11/2009	0.4519	0.3852
	08/2008 - 03/2009	0.5902	0.5330
unemployment			
	08/2008 - 12/2014	0.5774	0.4543
	08/2008 - 11/2009	0.5794	0.5126
	08/2008 - 03/2009	0.8285	0.7712
Buy and Hold			
	08/2008 - 12/2014	0.5735	0.4533
	08/2008 - 11/2009	-0.0866	-0.0906
	08/2008 - 03/2009	-0.3282	-0.3978

Table 4: Wikipedia & Dow 30 Strategy Returns

5 Findings and their Socioeconomic Implications

The final strategy we tested, which traded the DJIA on the basis of negative general economic indicators was the most successful and the most surprising. The following figures (2, 3, 4) illustrate the individual and cumulative geometric returns of the Wikipedia & Economic Terms strategy.

The implications for these findings are impressive: being able to utilize publicly available data to interpret and predict market moves can benefit more than just investors. During the sub-prime mortgage crisis of 2008-09/10, U.S. households lost over $7 trillion dollars in their total net worth.[17] Trading strategies such as these can help to not only insulate investors from down-side market risk that is experienced during bear markets and economic crises, but can also help to build predictive models that can alert regulatory agencies and governments of an impending economic crisis, by gauging the public sentiment through data from sources such as Wikipedia

[17] `http://money.cnn.com/2011/06/09/news/economy/household_wealth`

and Google Trends.

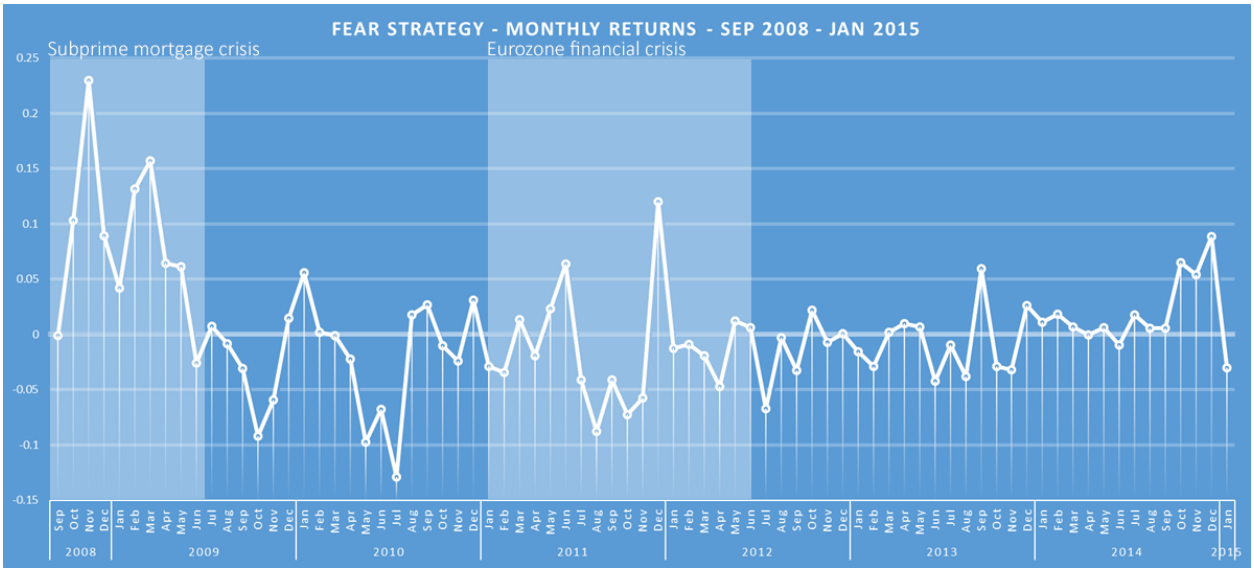

Figure 2: Monthly individual returns from fear strategy

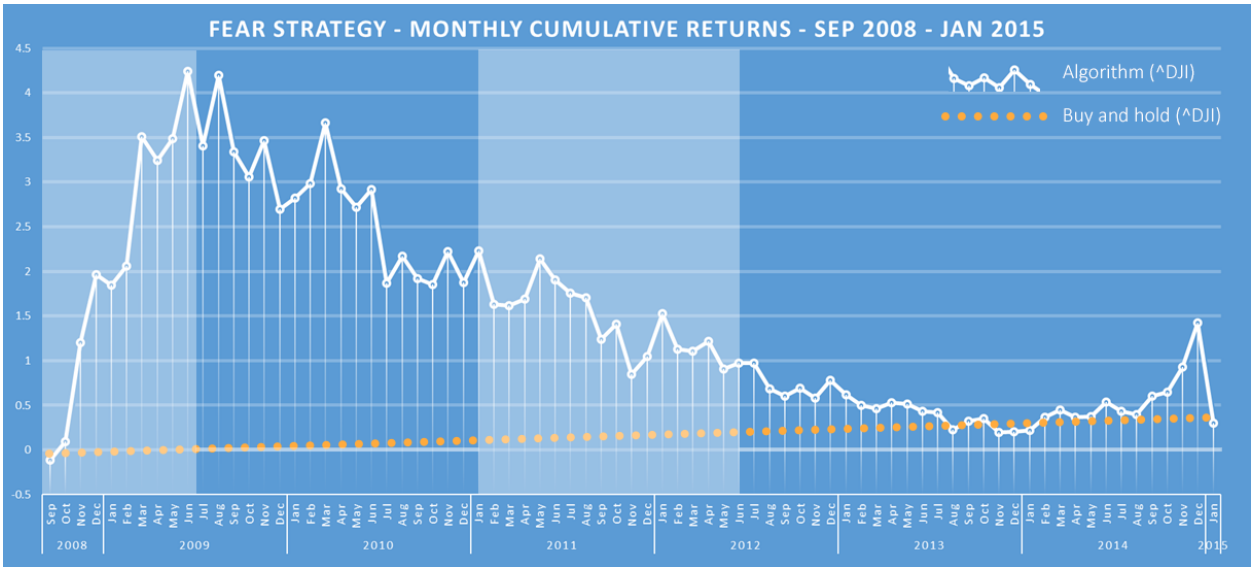

Figure 3: Monthly cumulative returns from fear strategy

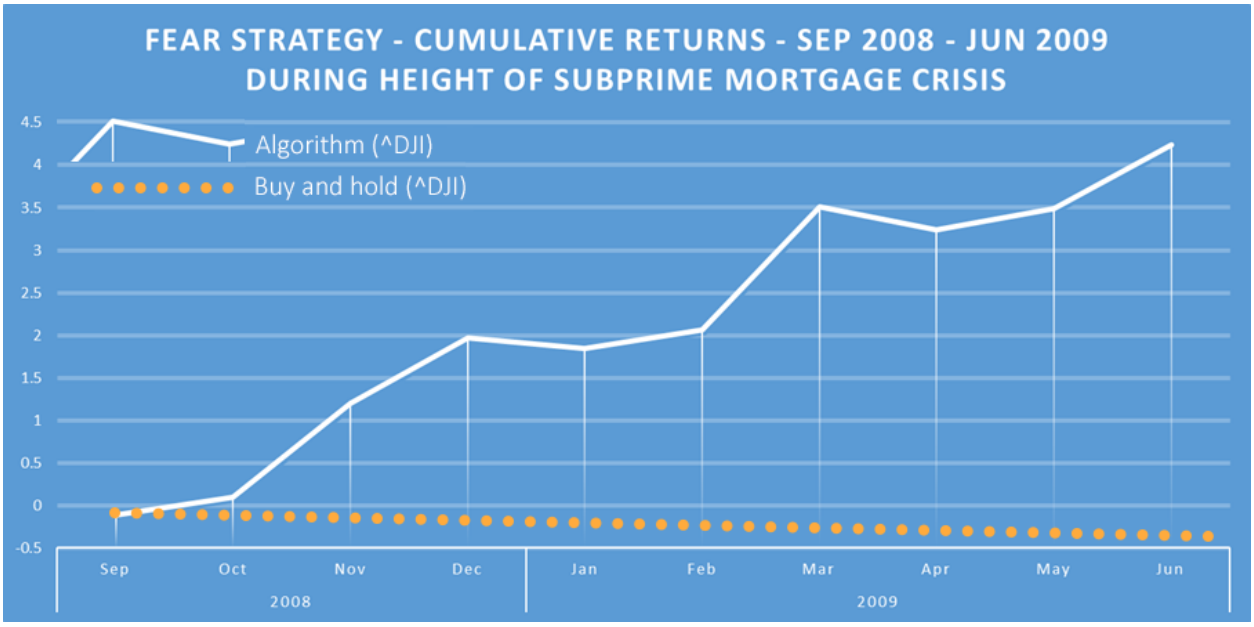

Figure 4: Monthly cumulative returns magnified during the subprime mortgage crisis

6 Conclusions

The most successful strategy was the final one: looking for connections between Wikipedia pages on negative general economic terms and market moves. We discovered that there can be a lot of error in the raw page view data that is provided. Indeed, it is likely that not a large amount of the page views are from people looking to invest. This makes it even harder for a potential strategy to be successful. If we had access to internal data through which we could categorize and filter page views by location, previous search history, and IP address, the quality of the data would be substantially improved and could allow us to develop more accurate models. However, it is nearly impossible to obtain this detailed data as it includes personally identifiable information.

Learning that many factors can have a hand in market and pricing fluctuations is undeniably eye-opening. Today, data is being collected at an ever increasing rate. With more comprehensive data, investor sentiment could become a new de facto standard in market analysis, and could potentially help to avoid massive downside market risk through actively measuring the market and investor sentiment.

The field of quantifying and researching investor sentiment is still in its infancy. Future research is vital in order to better understand the interplay between investors and the market. This study effectively built a technological framework for sentimental asset analysis, which I plan to expand on in the future in order to continue testing sentiment analysis techniques.

Appendix A Selected Code

The *adjust_for_holidays()* function is utilized by the trading function to check if the beginning of the week lies on a holiday or market holiday. It returns the dates corrected for any market holidays.

```
  def self.adjust_for_holidays(from, to)
    end_day = from + 14
      us_holidays = Holidays.between(from, end_day, :us)
      holidays = Array.new
      holidays.push(Date.civil(2010,7,5))
      holidays.push(Date.civil(2011,12,26))
      holidays.push(Date.civil(2012,1,2))
      holidays.push(Date.civil(2012,10,29))
      holidays.push(Date.civil(2012,10,30))
      us_holidays.each do |hd|
        holidays.push(hd.values[0])
      end

      if holidays.include? from
        loop do
          from = from + 1
          break unless holidays.include? from
        end
      end

      if holidays.include? to
        loop do
          to = to + 1
          break unless holidays.include? to
        end
      end
    return from, to
  end
```

The following code is a snippet from the trading function for the strategy that trades the DJIA based on view counts of the Dow 30 companies. This snippet iterates on each week, making a decision and calculating cumulative returns.

```
loop do
  differences = Array.new
  companies.each do |company|
    pages = Pageview.where("company_id = ? AND ((date <= ?) AND (date
        >= ?))", company.id, end_date, start_date).order("date ASC")
    # Get the pageview counts as a subset from original data for the
       delta t period
    views_d_wk = pages.where("((date <= ?) AND (date >= ?))",
       end_d_wk, start_d_wk)
    # Get the pageview counts as a subset from original data for the
       t week period (week before 'today')
```

```
  views_t_wk = pages.where("((date <= ?) AND (date >= ?))",
     end_t_wk, start_t_wk)

  # Sum the views and average them, then find the difference
  avg_d_wk = views_d_wk.sum(:views).to_d / views_d_wk.count
  avg_t_wk = views_t_wk.sum(:views).to_d / views_t_wk.count
  difference = avg_t_wk - avg_d_wk
  differences.push(difference)

  puts "#{company.name}: Average this week: #{avg_t_wk}"
  puts "#{company.name}: Average last week: #{avg_d_wk}"
  puts "#{company.name}: Difference: #{difference}"
end

difference = differences.sum
from, to = Company.adjust_for_holidays(current_day, current_day +
   7)


if difference > 0
  # Interest is increasing, time to get out.
  sh = StockHolding.new
  sh.date = from
  sh.ticker_id = Ticker.dow.id
  sh.decision_id = Decision.sell.id
  sh.shares = 1

  sh1 = StockHolding.new
  sh1.date = to
  sh1.ticker_id = Ticker.dow.id
  sh1.decision_id = Decision.buy.id
  sh1.shares = 1

  initial -= sh1.price
  cashed_out += sh.price
  profit += (sh.price - sh1.price)
  single_return = (sh.price - sh1.price) / sh1.price
  single_geo_return = Math.log(sh.price) - Math.log(sh1.price)
  total_geo_return += single_geo_return
  total_return += single_return
  puts "Sold (short) @ #{sh.price}, Bought @ #{sh1.price}"
  puts "Incremental gain: #{single_return.to_s}"

  decisions.push(["SELL", sh.date, sh.price.round(2).to_s, "BUY",
     sh1.date, sh1.price.round(2).to_s, (sh.price - sh1.price).
     round(2).to_s])
  oreturns.push([sh1.date, single_geo_return])
```

```
  elsif difference <= 0
    # Interest is decreasing, time to get in.
    sh = StockHolding.new
    sh.date = from
    sh.ticker_id = Ticker.dow.id
    sh.decision_id = Decision.buy.id
    sh.shares = 1

    sh1 = StockHolding.new
    sh1.date = to
    sh1.ticker_id = Ticker.dow.id
    sh1.decision_id = Decision.sell.id
    sh1.shares = 1

    initial -= sh.price
    cashed_out += sh1.price
    profit += (sh1.price - sh.price)
    single_return = (sh1.price - sh.price) / sh.price
    single_geo_return = Math.log(sh1.price) - Math.log(sh.price)
    total_geo_return += single_geo_return
    total_return += single_return
    puts "Bought @ #{sh.price}, Sold @ #{sh1.price}"
    puts "Incremental gain: #{single_return.to_s}"

    decisions.push(["BUY", sh.date, sh.price.round(2).to_s, "SELL",
       sh1.date, sh1.price.round(2).to_s, (sh1.price - sh.price).
       round(2).to_s])
    oreturns.push([sh1.date, single_geo_return])
  end
```

The *retrieve_pageviews_span()* function gathers pageview data that has been previously downloaded from Wikipedia.

```
  def retrieve_pageviews_span(beginYear, beginMonth, endYear,
     endMonth)
    beginDate = Date.parse("#{beginYear}-#{beginMonth}-1")
    endDate = Date.parse("#{endYear}-#{endMonth}-1")
    if beginDate > endDate
      puts "Error: End date must be after begin date."
      return
    else
      current = beginDate
      loop do
        Pageview.retrieve(current.year, current.month, self)
        break if current == endDate
        current = current + 1.month
      end
      return
    end
```

```
  end
```

The *create_dow_company()* function will import Yahoo Finance CSV data and create a ticker record and corresponding StockPrice data.

```
def self.create_dow_company(name, ticker, exchange)
   t = Ticker.new
   t.ticker = ticker
   t.exchange = exchange
   t.save

   c = Company.new
   c.name = name
   c.ticker_id = t.id
   c.save

   c.retrieve_pageviews_span(2008,1,2014,12)

   StockPrice.import_stock_prices(ticker)

   return
 end
```

The *get_stock_prices()* function uses the open-uri gem to download and save stock price data directly from the Yahoo Finance engine.

```
def self.get_stock_prices(ticker)
   download = open("http://ichart.yahoo.com/table.csv?s=#{ticker.
      ticker.upcase}")
   IO.copy_stream(download, "data/#{ticker.ticker.upcase}.csv")
end
```