Can we predict the financial markets based on Google's search queries?

Marcelo S. Perlin*
Departamento de Administração
Universidade Federal do Rio Grande do Sul

João F. Caldeira
Departamento de Economia
Universidade Federal do Rio Grande do Sul

André A. P. Santos
Departamento de Economia
Universidade Federal de Santa Catarina

Martin Pontuschka
Departamento de Administração
Universidade Federal do Rio Grande do Sul

Abstract

In this paper we look into the interaction of Google's search queries and several aspects of international equity markets. Using a novel methodology for selecting words and a VAR modeling approach, we study whether the search queries of finance related words can have an impact on market uncertainty, log returns and traded volume of four different english speaking countries. Our overall main results are positive, we find several words that have a robust impact over the explained variables. Particularly, the most robust result we find is that an increase of the search queries including the word *stock* predicts an increase of volatility and decrease of index prices in the next week. With the support of this finding, we investigate the performance of a market timing strategy based on the search frequency of this word and compare it against random words from the Word-Net database and a naive buy and hold strategy. The results of this empirical application are positive and particularly stronger during the global crisis of 2008, indicating the potential of this new set of data in predicting financial markets.

JEL: G10, G11, G15
keywords: market efficiency, market microstructure, investor attention, google trends

*Corresponding author (marcelo.perlin@ufrgs.br).

1 Introduction

The price oscillation of contracts in the financial markets is the result of the interaction of a large pool of participants. Highly capitalized financial institutions and individual investors share their financial views by trading according to their expectations. An important piece of the puzzle of how markets are organized is related to the way that economic agents behave when faced with different information. Academically, this is considered a black box since, for many different reasons, no reliable data can be gathered regarding the behavior of each individual investor. This leaves us with a large unexplored gap in the understanding of financial markets inner mechanisms as we only see the output of the interaction of the different agents in the form of prices and traded volumes, but never the individual and true mindset that drives these. From the academic side, the question of how fast information reaches the participants and affects their trading decisions has generated one of the pillars of financial theory, the market efficiency theory (Fama, 1965, 1970).

However, with the advance of technology we are experiencing a revolution in how social information is collected and used. The broad collective utilization of internet search web pages such as *Google*, *Yahoo*, *Bing* and others, offers a rich data that can be used to better understand systematic effects in the general population as the popularity of internet use increases. As a simple example, the frequency of which a particular region of the globe searches for flu symptoms can provide an estimate of the likelihood of a flu outbreak in that particular area (Dugas et al., 2012). Nonetheless, search frequency data has been applied to a range of topics, not only the prediction of diseases (Dugas et al., 2012; Ortiz et al., 2011) but also consumer behavior (Carriere-Swallow and Labbe, 2013), prediction of economic variables (Choi and Varian, 2012) among others.

Closer to the financial aspect of analyzing internet search queries, this data can be seen as a channel that allows access to systematic effects impacting market participants. While we can't see or measure the specific and individual behavior of investors, we can at least analyze systematic patterns in social data. If one observes an increase of search frequency for a particular word in period t, it could provide a signal of what will be the trading behavior of investors in $t + k$. The search frequency pattern might also indicate a systematic effect that could be unobservable in any other way. As an example, a decrease in the mood of the investors can certainly impact their trading decisions, which can also impact the frequency of search queries for certain words. Therefore, using search frequency data one can provide new ways to better understand the systematic effects in the population and their relationship to the dynamics of financial markets. Such approach is clearly interesting when considering the importance of financial markets to the economy and society in general.

Past studies that have looked at social media data and relate it to the financial markets have been relatively successful. In Bollen et al. (2011) Twitter messages are read by a mood tracking tool which distinguishes between calm, alert, sure, vital, kind and happy messages. After processing the data, the authors measure its relationship with DJIA closing values. The main finding of this work is that some of the mood intensities can help explain the variation in the market index.

Closer to the case of using search frequency data, Bordino et al. (2012) tests the relationship between trading volume and ticker search queries in *Yahoo* for a sample of US stocks. The main result is positive, meaning that an increase in the search frequency of stocks ticker can explain changes in the traded volume of the same stock. This work is similar to Da et al. (2011) however, the authors of this earlier study looks into several different variables, including a transformation of traded volume (turnover), also using lags in order to measure the forecasting ability of Google search data. The results are also positive where an increase in SVI (search volume index) indicates higher stock price within the next two weeks and also a larger first-day return and log run underperformance of IPO stocks. A similar study of this kind was performed in Dimpfl and Jank (2011) and Vozlyublennaia (2014) with similar results, however they used the names of stock market indexes (e.g. S&P500, NASDAQ) as the words from which they sampled the search frequency.

In the work of Preis et al. (2013) the performance of a trading strategy is tested by defining the trading rules with the search volume of particular words. The authors defined their list of words with the help of Google Sets, a tool that offers a list of words that have the highest semantic relationship to a group of expressions. In this case,the benchmark expressions were terms related to *stock market*, resulting in 98 words such as *debt*, *investment*, *bonds*, and so on. The paper reports that the strategy based on search frequency presented positive excessive returns over the benchmark, indicating the potential use of search frequency data as a trading tool. The authors also show that the word *debt* is the one with the best overall results.

In this paper we look for extending such a result related to the impact of internet search frequency over financial markets by investigating a larger set of words and a more extensive financial dataset. A critic to the previous studies is that they mostly used one country only as the basis of the study. This is not optimal since internet search queries are impacted by cultural differences among the countries and also by their financial

history. The word *debt* might be full of meaning to the American population given the debt crisis of 2009, but not so much to other countries that presented higher economic resilience to the episode. If the main objective of the studies is to find internet search queries that impact the financial markets in general, then they are clearly biased by using a single country. Therefore, their results can only be applied to a specific country and not the international financial markets as a whole. We overcome these issues by using four countries with highly capitalized stock markets in the research.

Another issue with the previous studies is related to the word selection used for gathering data of internet search frequencies. Most of the studies used ticker symbols, which is a narrow and limited set of expressions that may shadow stronger results for a more comprehensive dataset. In this research we follow the work of Preis et al. (2013) and do not limit ourselves to individual stocks tickers or market names, but to a larger sample of words that are related to finance. However, we innovate in the methodology for selecting the words by using a finance dictionary and four different finance related books. In our modeling approach we also improve the methodology by using a general VAR model and granger causality tests, which allows for a two way causality test of the variables in question, that is, we test not only the impact of search queries over the financial markets but also whether financial markets can impact the search volume of specific words.

Our main results are positive. We find that a significant portion of the chosen set of words is able to robustly affect different aspects of the financial market such as traded volume, volatility and log return. Among all countries, we pay special attention to the word *stock*. In this case, the results are robust across the different countries and show that an increase of the search queries including this word can predict an increase of volatility and a decrease of stock market index prices. This result suggests that investors execute search queries related to the word *stock* prior to a sell decision. The forecasting power of the search frequency of this specific word is tested in an empirical application with an out of sample framework. The results are also positive, even when comparing them to the results found for random words from the Word-Net database. This paper indicates the potential of using social data, in this case internet search volume, in financial research.

2 The Data

Our study uses five sources of data to reach its objectives: a finance dictionary, four finance related books, Google's search volume and financial data for each of the countries in the study: United States, United kingdom, Australia and Canada. We select these countries based on two datasets from the World Bank database: stock market capitalization and percentage of internet users. First we rank all countries from high to low according to both metrics and we build a new ranking based on the average of the rankings from the two previous indicators. This provides a new ranking that, when again sorted, presents a list of countries with high capitalized markets and with a high percentage of internet users, both of which are desirable qualities for our study. Once this list is built, we further refine the countries by selecting four of the top of the aggregate ranking where English is the main language.[1]

2.1 Selecting the words

One of the crucial aspects of this research is the choice of the words from which we gather the search query volume data. Following the objective of the study, it is natural that the choice of words must be biased towards finance. Therefore, selecting these words represents a crucial part of the study (Challet and Ayed, 2013). While it would be easy to choose arbitrary words such as *investments*, *portfolio* and *diversification*, there would be a clear loss of scientific objectivity since the choice of words was particular to the researchers.

In the work of Preis et al. (2013) the words were selected based on the output of Google Sets, a spreadsheet tool that provide semantically related words to a set of expressions. Our critic in using this procedure is twofold. First, the Google Set tool is no longer available as of September 2014, therefore one cannot replicate or use the same procedure in different scenarios. Second, while the use of Google Sets yields words that are related to the topic in question, it can also lead to bizarre choices such as the words *restaurant*, *cancer*, *movie* among others (Preis et al., 2013; Challet and Ayed, 2013).

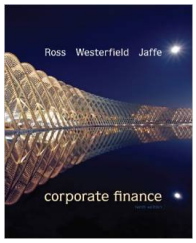
[1] A previous version of the research used data from the BRICS countries, with various national languages. The set of English words selected in the research was then translated to the specific local language using Google Translate. While this was an interesting approach, we found several problems with the resulting Google trends dataset because of the unavailability of the data. Most of the selected translated words had low search volume which resulted in zero entries from Google Trends. Given this issue, we decided to work only with countries with developed equity markets, a large proportion of internet users and where English is the main spoken language.

In this paper we innovate by using a finance online dictionary and four different finance textbooks to weight how strong is the relationship of each expression to the topic of finance. The first step was to extract all words from the internet Finance dictionary of Investopedia[2] as our benchmark of words that are related to finance. This primary dataset is composed of 14.479 unique sets of finance expressions (one of more words) such as *absolute interest*, *safe haven*, *Municipal Bond* and many others.

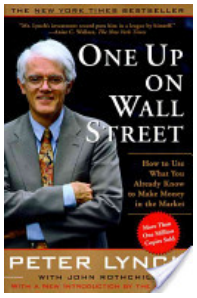
The second step is to use the previous set of words and to count the number of times that each term in the Investopedia dataset was found in the four finance books. We diversify the choice of the books by choosing two popular academic textbooks and two others that, at the time, were the two highest selling books in the section of finance from Amazon[3]. Next Figure 1, we show the cover page of each book and their reference.

Figure 1: Books used to find the set of words

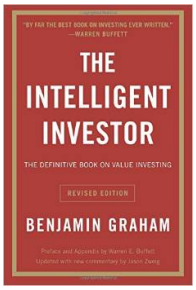
(a) Jaffe et al. (2005)



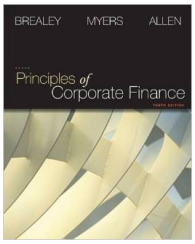
(b) Lynch (2000)



(c) Graham and McGowan (2005)



(d) Brealey (2011)



The intuition in using a set of textbooks is that it contains text for the specific field of finance, therefore we let the books to imply the set of words that are suitable candidates to the research in question. Next, Table 1, we show the selected fifteen English words, along with their total number of occurrences in the four books. As expected, we see the words *Finance*, *Capital* and *Value* at the list which also includes the word *Debt*, previously used in Preis et al. (2013).

2.2 Google's Search Volume (Google Trends)

Recently *Google* provided to the general public free access to a tool called *Google Trends*. Given an expression and a geological location (country), this website provides information of the internet search frequency related to both inputs. If there is sufficient search volume, the data is provided in the weekly frequency with a relative (normalized) structure, where the values range from 0 to 100. In order to reach these relative values, each nominal search volume for a particular period (week) is divided by the total search volume in the requested period. After that, the data is normalized so that the maximum value is 100 and the minimum is 0 (Choi and Varian, 2012).

Google trends calculates search frequency based on all uses of the word. As an example, the search frequency for the word *bread* will also include *white bread*, *brown bread* and so on. This means that looking at the search volume for a particular word one can find the search frequency for many variations that use the

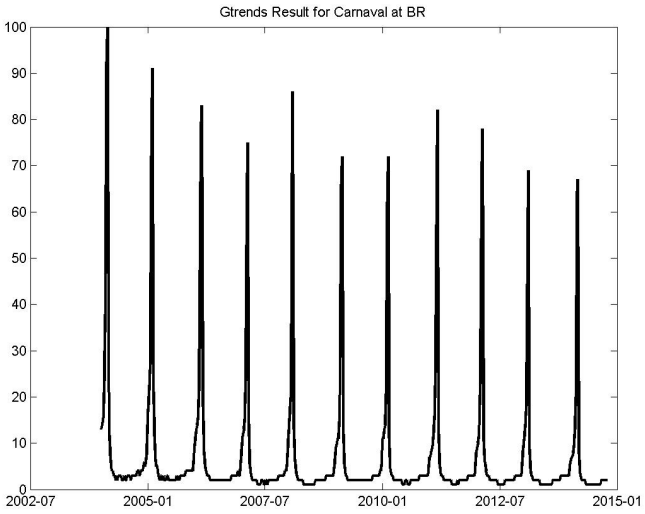
[2] http://www.investopedia.com/dictionary/
[3] www.amazon.com

Table 1: The fifteen words selected for the research

Number	Selected expression	Number of occurrences
1	Finance	2078
2	Cap	1752
3	Capital	1715
4	Corporate Finance	1355
5	Value	1298
6	Par	1077
7	Stock	965
8	Market	871
9	Risk	832
10	Cash	745
11	Dividend	739
12	Journal	718
13	Option	605
14	Year	547
15	Debt	540

same word. This is interesting from the research side since it allows for a diversified set of data. Also, words with low search volumes and repeated searches from the same user don't count towards the search frequency index of Google Trends. As an illustration of how the data is presented, next we provide a time series of Google search volume for the word *Carnival* in Brazil. [4]

Figure 2: An example of Google Trends data



In Figure 2 we can see that the Google Trends data has a maximum of 100 and the rest of the values are based on this maximum frequency. We also see that the time series data has a strong seasonality in the early months of each year. The pattern is easily justified since Carnival, the popular Brazilian festival, usually begins in February, meaning that there will be a significant number of search queries for this term around the holiday. In the econometric models described latter the seasonality of the Google trends data will be dealt with in order to isolate the particular effect we are investigating.

[4] In the corresponding author's personal website one can find and download the Matlab code that imports Google Trends data directly into the workspace. This tool allowed for a large scale download of Google Trends data and it proved to be very useful for this type of study.

2.3 Financial Data

The financial data used in the study is composed of stock market index prices and the traded volume of its constituents, weighted by the individual asset's participation. The prices of the indexes are used to calculate return and volatility. All of the financial data is measured daily and, following the frequency of the Google trends data, we aggregate it into the weekly frequency. The actual equations for aggregating the financial variables are given next:

$$Volat_t = \sqrt{\frac{\sum_{j=1}^{nDays_t}(R_i - E(R_i))^2}{nDays_t}} \tag{1}$$

$$Ret_t = \frac{\sum_{j=1}^{nDays_t} R_i}{nDays_t} \tag{2}$$

$$Vol_t = \frac{nDays_t^{-1}\sum_{j=1}^{nDays_t} Vol_i}{10000} \tag{3}$$

For Equations 1, 2 and 3, variable R_i represents the log daily return for each day within the week t and $nDays_t$ is the number of days for that particular week. Notice that for each of the equations we use an average that weights by the number of days within a week, that is, we are considering the effect of weeks where the number of trading days is different than five, possibly caused by the closure of the market.

The variables built using Equations 1, 2 and 3 are used as dependent variables in the econometric models, therefore we approach the impact of search frequency over three dimensions of market behavior: return, volatility and traded volume.

Table 2: Descriptive statistics for financial data of the four countries in the dataset

Country	Market Index	Mean of Weekly Mean Log Ret	Std Dev of Weekly Mean Log Ret	Mean of weekly Traded Volume
US	SP500	0,026%	0,50%	3704557,929
UK	FTSE	0,010%	0,47%	1132023,27
AU	S&P/ASX 200	0,007%	0,43%	942331,557
CAN	S&P/TSX	0,038%	0,45%	181527,1598

Table 2 report some basic statistics for the trading data of each country. The American and the Canadian equity markets presented the highest bull phase during the period, with the highest average log returns. The volatility, however, is fairly comparable in between the markets, with a minimum of $0,43\%$ for the Australian market index and a maximum of $0,5\%$ for the American market .

3 The Models

The econometric models used in the research have the purpose of measuring the endogenous effect of Google trends and its predictive power over financial markets. We use a structural VAR as our main model that will provide insights regarding the endogenous relationship of search frequency over the dependent variables. This means that we not only test for the effect of a change of search frequency of certain words in Google but also the inverse, that is, the effect that the financial markets can have regarding the volume of particular search queries. For each of the models we interact three different dependent variables: difference of volatility ($\Delta Volat_t$), return (R_t) and difference of traded volume (ΔVol_t).

$$y_t = \alpha_1 + \sum_{p=1}^{OptLag} \beta_p y_{t-p} + \sum_{p=1}^{OptLag} \lambda_p \Delta GTrends^*_{t-p} + \epsilon_{1,t} \tag{4}$$

$$\Delta GTrends^*_t = \alpha_2 + \sum_{p=1}^{OptLag} \gamma_p \Delta GTrends^*_{t-p} + \sum_{p=1}^{OptLag} \phi_p y_{t-p} + \epsilon_{2,t} \tag{5}$$

In the system of equations of 4 and 5, the variable y_t is a placeholder for $\Delta Volat_t$, R_t and ΔVol_t. The variable $GTrends_t^*$ is the original Google trends data without the seasonal effect. We follow McTier et al. (2013) and define $GTrends_t^*$ as the residual from the regression $GTrends_t = \alpha + \sum_{k=1}^{11} \phi_k D_{k,t} + \epsilon_t$, where the dummy $D_{k,t}$ takes value one if date t is in month k (1..12) and zero otherwise. We run the VAR model for each country and each word in the list of Table 1. The lag of the system ($OptLag$) is determined in a dynamic fashion, using the sequential LR test as described in Lütkepohl (2007). Additionally, we perform two-way Granger Causality tests using the VAR model, which will indicate how strong the financial data can predict Google queries and vice versa.

4 Results

We start our analysis by presenting the estimation results of the VAR model in Tables 3, 4 and 5. As a rule of thumb, we define a result as robust if the analyzed parameter presents statistical significance for at least three out of the four countries, and the same sign in all cases.

Table 3: **Estimation results for the VAR model using volatility ($\Delta Volat_t$)**

The table reports the estimation results for the following VAR model:

$$\Delta Volat_t = \alpha + \sum_{p=1}^{OptLag} \beta_p \Delta Volat_{t-p} + \sum_{p=1}^{OptLag} \lambda_p \Delta GTrends^*_{t-p} + \epsilon_{1,t}$$

$$\Delta GTrends^*_t = \alpha + \sum_{p=1}^{OptLag} \beta_p \Delta GTrends^*_{t-p} + \sum_{p=1}^{OptLag} \phi_p \Delta Volat_{t-p} + \epsilon_{2,t}$$

The statistical analysis in the second and third column within each country tests the null hypothesis that $\sum_{p=1}^{OptLag} \lambda_p = 0$ and $\sum_{p=1}^{OptLag} \phi_p = 0$, respectively. The symbol *, **, *** represents significant p-values at the 10%, 5% and 1% level.

	USA			UK			AUS			CAN		
Word	Optimal Lag	Sum of λ_p	Sum of ϕ_p	Optimal Lag	Sum of λ_p	Sum of ϕ_p	Optimal Lag	Sum of λ_p	Sum of ϕ_p	Optimal Lag	Sum of λ_p	Sum of ϕ_p
Finance	5	1.38**	-0.03	4	0.29	-0.03	5	0.44***	0.25***	5	0.64***	0.20**
Cap	5	0.58	-0.08	5	-0.61	-0.05	5	-0.46	-0.12	5	-0.31	0.01
Capital	5	0.39	0.01	5	-0.01	-0.01	5	0.20	0.15	5	0.58*	-0.02
Corporate Finance	5	0.02	-0.04	5	-0.25	0.12	5	0.09	-0.57	5	-0.10	-0.14
Value	5	0.35	0.06	4	0.19	0.03	5	0.35	0.25**	5	0.10	0.07
Par	5	0.16	0.02	5	0.06	0.07	5	0.10	-0.17	5	0.20	0.12
Stock	5	0.91*	-0.06	5	0.73**	-0.25***	5	1.41***	0.02	5	0.58***	0.06
Market	5	1.15***	0.01	5	-0.22	-0.11	5	0.48	-0.05	5	0.79*	0.08
Risk	5	0.08	0.11	4	-0.20	-0.02	5	0.00	0.25	5	0.17	0.01
Cash	5	-0.30	0.04	4	1.27***	0.02	5	0.24	0.16	5	0.14	-0.08
Dividend	5	0.11	0.14	5	-0.22	0.08	5	0.20	0.39***	5	0.13**	0.20
Journal	5	0.12*	0.05	4	0.06	-0.00	5	0.28**	0.12*	5	0.47	0.10
Option	5	-0.49	-0.00	4	-0.03	-0.00	5	0.05*	0.15	5	-0.07	0.10
Year	5	-0.26	-0.17	4	-0.04	-0.13	5	-0.09	0.10	5	-0.15	-0.23
Debt	5	1.37***	-0.10	5	0.31	-0.10	5	-0.23	0.01	5	0.64*	-0.17

Table 4: **Estimation results for the VAR model using log returns (R_t)**

The table reports the estimation results for the following VAR model:

$$R_t = \alpha + \sum_{p=1}^{OptLag} \beta_p R_{t-p} + \sum_{p=1}^{OptLag} \lambda_p \Delta GTrends^*_{t-p} + \epsilon_{1,t}$$

$$\Delta GTrends^*_t = \alpha + \sum_{p=1}^{OptLag} \beta_p \Delta GTrends^*_{t-p} + \sum_{p=1}^{OptLag} \phi_p R_{t-p} + \epsilon_{2,t}$$

The statistical analysis in the second and third column within each country tests the null hypothesis that $\sum_{p=1}^{OptLag} \lambda_p = 0$ and $\sum_{p=1}^{OptLag} \phi_p = 0$, respectively. The symbol *, **, *** represents significant p-values at the 10%, 5% and 1% level.

	USA			UK			AUS			CAN		
Word	Optimal Lag	Sum of λ_p	Sum of ϕ_p	Optimal Lag	Sum of λ_p	Sum of ϕ_p	Optimal Lag	Sum of λ_p	Sum of ϕ_p	Optimal Lag	Sum of λ_p	Sum of ϕ_p
Finance	3	-0.22***	0.06***	3	-0.09	-0.01	5	-0.29**	0.27**	5	-0.10**	0.33**
Cap	4	-0.03	0.04	5	-0.17	0.04	5	0.01	-0.09	3	-0.06	-0.08
Capital	5	0.01	-0.06***	5	-0.03	0.08	3	-0.09	-0.03	5	-0.25	0.17
Corporate Finance	5	0.03	0.28	5	-0.01	0.17	3	0.02	0.58*	3	0.00	0.13
Value	4	0.02	0.00**	5	-0.11	0.15	2	-0.01	0.08	2	0.02*	0.03
Par	4	-0.02	0.05	5	-0.14	0.12	3	0.02	0.06	4	-0.27*	0.03
Stock	3	-0.23***	0.18***	5	-0.02***	0.34**	4	-0.28***	0.26*	5	-0.03***	0.29***
Market	5	-0.46*	0.19	5	-0.13	0.13	3	-0.17	0.12	3	-0.25*	0.31***
Risk	4	0.05	-0.01***	4	-0.04	0.05	2	-0.03	0.03	4	-0.07	0.10*
Cash	5	0.13	-0.08	4	-0.31	0.18*	4	-0.17	-0.04	5	0.12	-0.07
Dividend	4	-0.00	-0.10	5	0.02	0.08	4	0.02	-0.11	5	0.04	0.08
Journal	5	0.11	-0.02*	4	-0.15	0.03	2	-0.04	0.04	3	-0.04	0.04
Option	4	0.11	-0.03	5	-0.04	0.10	5	-0.11	0.17	4	-0.04	0.07
Year	4	0.07	-0.06**	4	0.05	-0.02	5	0.06	-0.17	4	0.09	-0.06
Debt	5	-0.17	0.01	5	-0.19*	0.15	4	-0.23*	0.21	5	-0.32***	0.15

Table 5: **Estimation results for the VAR model using volume (ΔVol_t)**

The table reports the estimation results for the following VAR model:

$$\Delta Vol_t = \alpha + \sum_{p=1}^{OptLag} \beta_p \Delta Vol_{t-p} + \sum_{p=1}^{OptLag} \lambda_p \Delta GTrends^*_{t-p} + \epsilon_{1,t}$$

$$\Delta GTrends^*_t = \alpha + \sum_{p=1}^{OptLag} \beta_p \Delta GTrends^*_{t-p} + \sum_{p=1}^{OptLag} \phi_p \Delta Vol_{t-p} + \epsilon_{2,t}$$

The statistical analysis in the second and third column within each country tests the null hypothesis that $\sum_{p=1}^{OptLag} \lambda_p = 0$ and $\sum_{p=1}^{OptLag} \phi_p = 0$, respectively. The symbol *, **, *** represents significant p-values at the 10%, 5% and 1% level.

	USA			UK			AUS			CAN		
Word	Optimal Lag	Sum of λ_p	Sum of ϕ_p	Optimal Lag	Sum of λ_p	Sum of ϕ_p	Optimal Lag	Sum of λ_p	Sum of ϕ_p	Optimal Lag	Sum of λ_p	Sum of ϕ_p
Finance	5	-2.64	0.00	5	-0.30*	-0.08***	5	-0.10	-0.13***	5	-0.18***	-0.57***
Cap	5	-2.94*	-0.01	5	-0.76	-0.02	5	0.38	-0.08	5	-0.13	0.24
Capital	5	-4.34***	0.00	5	-2.10**	-0.01	5	-0.43	-0.10**	5	0.13	-0.80***
Corporate Finance	5	0.52	-0.04	5	-0.65	0.11***	5	0.06	-0.17	5	-0.03	-0.86***
Value	5	0.95***	-0.02*	5	0.68***	-0.04***	5	-0.06	0.02	5	-0.16***	-1.01***
Par	5	1.06	0.01	5	0.78	0.01*	5	-0.07	0.08	5	0.05***	-0.38***
Stock	5	-0.50	0.02**	5	-1.40	-0.10***	5	-0.59	-0.00	5	-0.07**	-0.25
Market	5	-2.24***	0.03***	5	-1.52***	-0.06***	5	-0.49	0.14**	5	-0.17	0.17
Risk	5	-1.38***	-0.03*	5	-0.98***	-0.12***	5	-0.39*	-0.09**	5	-0.06	-0.82***
Cash	5	2.30**	-0.02	5	1.19	0.05	5	-0.44	0.03	5	0.10	-0.36
Dividend	5	-1.05***	0.01*	5	-0.51**	-0.09	5	-0.34	0.05	5	-0.00	0.29
Journal	5	-1.64***	-0.02*	5	-2.05***	-0.05***	5	-0.26	-0.03	5	-0.20***	-0.33***
Option	5	-4.86	-0.01	5	-1.35*	0.02*	5	-0.38	-0.11	5	-0.08	-0.36
Year	5	4.62**	-0.00***	5	3.26***	-0.07***	5	0.95	0.18***	5	0.23	0.64***
Debt	5	1.80*	-0.02*	5	0.51***	-0.08***	5	-0.06**	-0.11***	5	0.29***	-0.63**

In Tables 3, 4 and 5 we show the estimation results for Equations 4 and 5. Notice that we report only the sum of ϕ_p or λ_p and not the parameters individually. The idea is to capture the long term dependence of Google trends data over the dependent variables and not the individual lag structure. The statistical significance of the sum of coefficients is calculated using a two way granger causality test, that is, we test the null hypothesis that the lag parameters of interest, ϕ_p or λ_p, are equal to zero in each VAR model.

From Table 3 we can see that two words presented a robust result for the case of the sum of λ_p, which we assume is the same sign of the sum of coefficients and at least three cases with statistical significance lower than 10%. These words are *stock* and *finance*, with positive values of sum of λ_p. This result shows that an increase of search frequency in this set of words has the potential of predicting a future increase of volatility in the different equity markets. For the results where volatility granger causes the frequency of search queries (sum of λ), we do not find a robust relationship in the data.

For the case of search frequency granger causing log returns of market indexes, Table 4, we again find robust results for the words *finance*, *stock* and *debt*. However, these are all negative sum of λ_p, meaning that an increase of search frequency of the set of words can impact future changes in the price index in a negative way. This suggests that the search queries of the selected words are mostly used before a systematic sell decision from investors, which in turn will decrease stocks prices. It is also interesting to note that the word *debt*, when used in a lagged model, goes the opposite way of the result in Preis et al. (2013), presenting a statistically significant negative sum of coefficients, with the exception of the US market where the null hypothesis was not rejected.

For the case of log returns granger causing search frequency, the only robust result we find is for the word *stock*, meaning that an increase of the market index price can predict a higher search frequency for this word. One could explain this result as a trend-following strategy, where investors search the internet before entering the market once they see a previous systematic positive jump in assets prices.

When looking at the results for volume, Table 5, we see that the results are in general stronger than with the previous two tables, with a higher density of significant coefficients. For the case of search frequency granger causing volume, we find that the words *journal* and *risk* have consistency, with negative and significant results in at least three of the four countries. The results for the test of market volume granger causing the search frequency in Google, we find the words *debt*, *journal* and *risk* with a robust negative effect of volume granger causing the search frequency.

The broad results of this econometric exercise are positive. Using a large financial database and search queries data, we find robust statistical relationships that suggest that the internet search queries have the potential to impact and serve as predictive indicators of different aspects of financial markets. The specific result from this research is the case of the word *stock*. Across all models and countries, the word *stock* presented the most statistical significance. We find that an increase of search queries related to *stock* in week $t-p$ can predict an increase of volatility and a decrease of stock prices in the following week t. We also find that a positive return of the stock market indexes can predict an increase of search queries related to the word *stock*, suggesting that investors lookup this specific word once they see a big jump in financial indexes.

5 An empirical application

In this section we explore the results from the econometric models by evaluating its predictive performance in a portfolio trading strategy. For this purpose, we use the search frequency of the word *stock*, which is the word with the most robust results in the previous tables. Our empirical application is based in the procedure described in Christoffersen and Diebold (2006) which uses forecasts of returns and volatility as input for building out of sample trading signals in a market timing strategy.

The idea of Christoffersen and Diebold (2006) is to create a trading sign I_t based on expected return and expected volatility, $I_t = F\left(\frac{\hat{\mu_t}}{\hat{\sigma_t}}\right)$ where $F(x)$ is the logistic function ($F(x) = \frac{exp(x)}{1+exp(x)}$) and $\hat{\mu_t}$ and $\hat{\sigma_t}$ are predictions of returns and volatility for time t, both calculated based on the VAR model of the previous section, Equations 4 and 5.

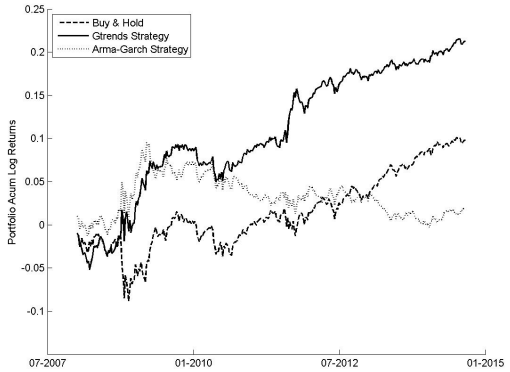
In order to test the out-of-sample predictability of the model, we divide the dataset into two smaller sub periods: the in-sample period starting from 01/01/2005 until 01/01/2008 and the out of sample period from 02/01/2008 until the last day of the data, 01/01/2014. As usual, we use the first sub-period to estimate $\Delta GTrends_t^*$ and the VAR model of Equation 4 using returns (R_t) and the difference of volatility ($\Delta Volat_t$) as the dependent variables. Based on this models we create predictions of expected return ($\hat{\mu_t}$) and contemporaneous volatility ($\hat{\sigma_t} = \hat{Volat_t}$) which are used as input in $\hat{I_t} = F\left(\frac{\hat{\mu_t}}{\hat{\sigma_t}}\right)$. Our simple strategy is to take

a long position in the index when $\hat{I}_t > 0.5$ and a short position otherwise. Notice that the performance of the strategy is a function of the predictive performance of econometric model. If the model predicts well the returns and the volatility, the resulting portfolio should have a positive return in a bull or a bear market.

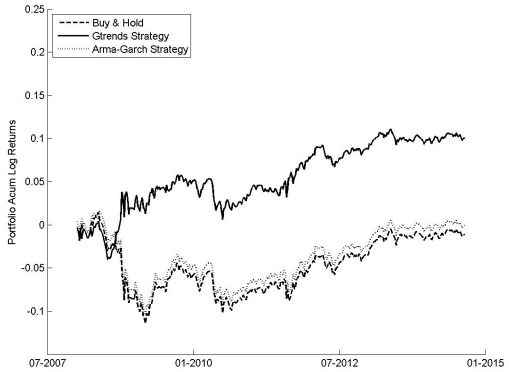
We compare the trading strategy to two benchmarks, a simple buy&hold strategy, that is, we buy the stock index at the beginning of the time period and sell it at the end, and also a strategy based on a ARMA(1,1)-GARCH(1,1) model. The idea of using the last one is to compare whether using the internet search frequency data one can have a better forecast than a simpler time series model. Next, Figure 3 we provide the accumulated return of these portfolios over time as an illustration of the results.

Figure 3: Accumulated log returns of the trading strategies

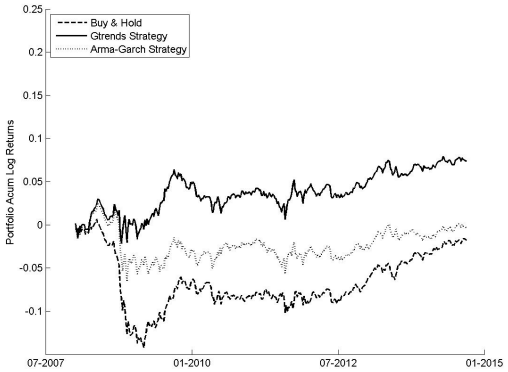
(a) USA - S&P500



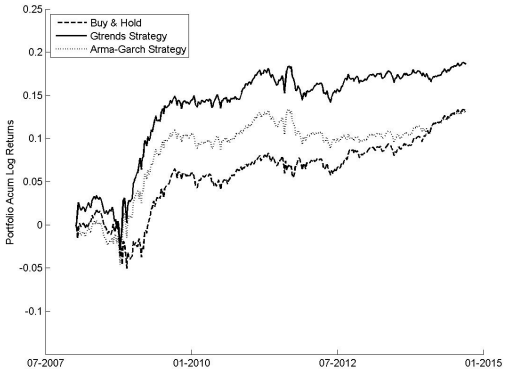
(b) UK - FTSE



(c) AUS - S&P/ASX 200



(d) CAN - S&P/TSX



From Figure 3 one can see that for all of the markets the trading strategy using Google trends data presented higher accumulated return than both benchmarks, the buy&hold strategy and the strategy using the ARMA(1,1)-Garch(1,1) model. A simple visual inspection already indicates the out of sample predictive power that the internet search frequency for the word *stock* has over financial markets returns and volatility. It is also interesting to see that the strategy performed well during the 2009 financial crises.

In this section we also compare the results of the trading strategy based on search frequency of the word *stock* against a dataset of words selected randomly. This robustness procedure described in Challet and Ayed (2013) tests how special (or significant) are the results with the word in question when comparing against the results found from a large set of words that have no significant meaning in Finance. In order to select the words, we use the lexical database of Word-Net[5] (Fellbaum, 1998). This database provides semantic classification of all words in the English language, including their conceptual meaning. Within Word-Net we search for all nouns and classify them based on their lexical domain[6], which represents the topic (or meaning)

[5] http://wordnet.princeton.edu/

[6] Column *lexdomain* in the MySQL/SQlite table/view *dict* of Word-Net.

that the noun belong to[7]. Within each group we select random words that compose 10% of the total. This procedures builds a diversified set of words, making sure that each sense id (or meaning) is represented. In total, we select 14.639 random words from Word-Net[8].

Once the set of random words is built, we proceed to download Google trends data associated with each. However, the data in Google trends is not guaranteed to exist since a low search volume for a particular word will not be recorded. Also, the Google trends data might be in the monthly frequency if the word does not have a significant search volume and this is undesirable since the trading strategy was build using weekly data. In order to respect the framework of the research, we control it by ignoring cases where the frequency is weekly and also the cases where the number of zero entries is higher than 50% of the total observations. In the end, we are left with 6510 cases of internet search frequency data of random words for USA, 4154 for UK, 2876 for AUS and 3564 for CAN. [9].

Based on the Google trends data of the selected random words, we proceed to repeat the same methodology we use for the word *stock* by building portfolios with the predictions of the VAR model, which is re-estimated for each new word. In order to assess the significance of the results from the word *stock* we simply calculate the proportion of cases from the random words that presents a historical portfolio with higher sharpe ratio than when using the word *stock*. Next, Table 6 we present this result along with other statistics of performance.

Table 6: Performance of the trading strategies

	US	GB	AU	CA
Total Return				
Buy & Hold	9,80%	-1,10%	-1,75%	13,20%
Gtrends (*stock*)	21,27%	10,06%	7,38%	18,63%
Arma-Garch	2,00%	-0,10%	-0,32%	13,04%
Risk (volatility)				
Buy & Hold	0,56%	0,52%	0,48%	0,50%
Gtrends (*stock*)	0,56%	0,52%	0,48%	0,50%
Arma-Garch	0,56%	0,52%	0,48%	0,50%
Sharpe Index				
Buy & Hold	0,050	-0,006	-0,010	0,075
Gtrends (*stock*)	0,109	0,056	0,044	0,107
Arma-Garch	0,010	-0,001	-0,002	0,074
% of portfolios with lower sharpe ratios				
Buy & Hold	0.58	0.11	0.46	0.29
Gtrends (*stock*)	0.98	0.79	0.97	0.63
Arma-Garch	0.31	0.14	0.62	0.28

The numerical results found in Table 6 leads to the same conclusions as the visual inspection of Figure 3. We can see that, for all countries in the study, the trading strategy using search frequency data presents higher total return and higher sharpe ratio than the other benchmark strategies. Notice that the risk of all strategies are the same. This is expected since using a threshold of 50% for I_t we force the strategy to trade everyday with a long or short position. That is, the difference in the vector of returns from one strategy to the next is simply the sign of the return of the index.

When looking at the results of comparing the performance (sharpe ratio) of random words against the word *stock*, we again see that the results are significant. For all countries, the percentage of random words that yields lower sharpe ratios averages at approximately 85%, with a maximum of 98% for the US and a minimum of 63% for CA. This clearly indicates that the word *stock* is indeed special and one cannot easily replicate its results with another word chosen randomly.

A further inspection of the trading results, however, adds an interesting information to the analysis. In Figure 4 we look at the variation of difference of performance between the strategy using Google trends data and the benchmark alternative (buy&hold). We also add a second axis to the figure with the values

[7]In appendix A one can find a table with the number of words for each sense ID (lexical domain) and also the number of words selected within each subgroup

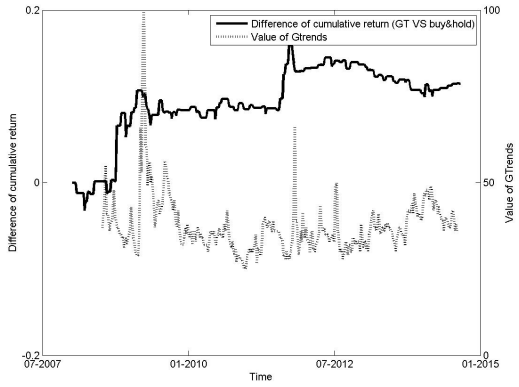
[8]The list of random expressions from Word-Net are not presented in the paper given its extension. However, it can be sent upon request.

[9]In Appendix B we present the results from handling the random words dataset from Google trends.

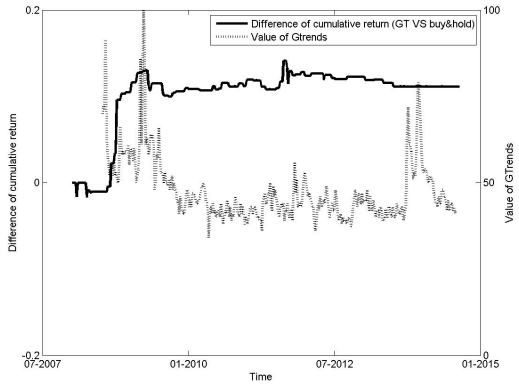
from Google trends. The idea is to look for consistency in the difference of cumulative returns and how the search frequency for the word *stock* could impact the performance of the out of sample trading strategy. If the strategy was robust, it should provide an upward slope of difference of returns, meaning that it was consistently able to beat the alternative. However, our results do not show such a consistency.

Figure 4: A comparative analysis of the difference of accumulated log returns (GTrends VS buy&hold) and values of search frequency

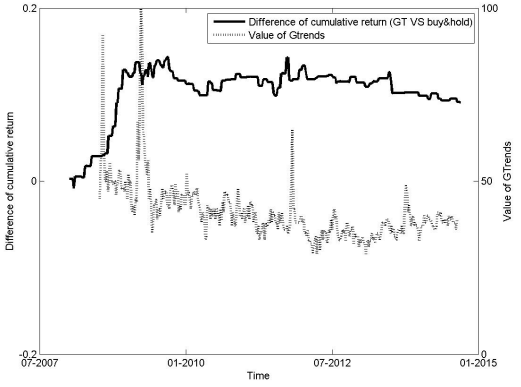
(a) USA - S&P500



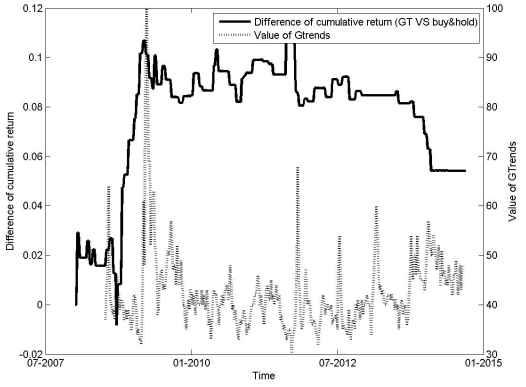
(b) UK - FTSE



(c) AUS - S&P/ASX 200



(d) CAN - S&P/TSX



We see from Figure 4 that for all countries, the biggest difference in accumulated returns happens at the 2009 and, after that, the difference is more stable, with exception of USA and CAN. We also see that at this time the level of search frequency for the word *stock* is also higher, reaching its maximum of 100 in all countries. This suggests that both results are related to the 2009 global financial crisis which has hit all markets in the sample. As an example for the impact in the US, the two-month period of January and February of 2009 represented the worst start to a year in the history of the $S\&P500$ with a 18.62% decrease in the stock price index. Needless to say that investors had the right motivations to be searching heavily for the word *stock* in this period.

This result shows that the predictability of Google trends towards stock market index return is stronger in the episode of 2009's financial crisis. We see that, in this time period, the trading strategy based on the predictability of the econometric model performed the best. This result suggests that the search frequency data can be specially helpful in predicting stock markets in episodes of systematic financial crisis. The recovery of stock prices after the crisis can be related to the decrease of the search frequency of the word *stock* and therefore a potential application would be to forecast a subsequent recovery of the stock market based on search frequency data.

6 Conclusion

In this paper we studied the impact of the search frequency of finance related words towards three different aspects of international financial markets indexes: volatility, log returns and traded volume. We innovate in this research in terms of scale: we use a large dataset of three different dependent variables and four english speaking countries. Using a robust approach and comparing the models across the different countries, our results show that social data, in this case internet search queries, do have predictive power over different aspects of financial markets, which corroborates with the general results found in the previous studies (Preis et al., 2013; Vozlyublennaia, 2014; Bollen et al., 2011; Bordino et al., 2012).

In the granger causality tests we find several robust results for different expressions where one in particular stands out, the word *stock*. Our models show that an increase of search frequency for this word in the previous weeks can impact future volatility positively and future returns negatively. This suggests that investors search for queries related to *stock* in the weeks preceding a large and negative jump in international stock market indexes.

In the empirical application of the results we build portfolios based on the out of sample predictions from the VAR model and again we find positive indications. The strategy which uses search frequency of the word *stock* presents higher return than any of the benchmark strategies, a naive buy&hold strategy and a strategy based on a Arma-Garch model. When comparing the results against 14.639 words selected randomly from Word-Net, again we find that the results for the word *stock* are particular and cannot be easily replicated. Interestingly, we also find that the predictability of the forecasting econometric model is stronger during the 2009's financial crisis, which indicates that this data might be particularly helpful in predicting systematic crashes and recoveries of the financial markets. However, since we can only go as far as 2004 with Google trends data, this hypothesis might need more time and new global crisis in order to be properly tested. Needless to say that this research exercise show the potential of using search query data in the study of financial markets.

References

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

Bordino, I., Battiston, S., Caldarelli, G., Cristelli, M., Ukkonen, A., and Weber, I. (2012). Web search queries can predict stock market volumes. *PloS one*, 7(7):e40014.

Brealey, R. A. (2011). *Principles of corporate finance*, volume 10. Tata McGraw-Hill Education.

Carriere-Swallow, Y. and Labbe, F. (2013). Nowcasting with google trends in an emerging market. *Journal of Forecasting*, 32(4).

Challet, D. and Ayed, A. B. H. (2013). Predicting financial markets with Google Trends and not so random keywords. *ArXiv e-prints*.

Choi, H. and Varian, H. (2012). Predicting the present with google trends. *Economic Record*, 88(s1):2.

Christoffersen, P. F. and Diebold, F. X. (2006). Financial asset returns, direction-of-change forecasting, and volatility dynamics. *Management Science*, 52(8):1273–1287.

Da, Z., Engelberg, J., and Gao, P. (2011). In search of attention. *The Journal of Finance*, 66(5):1461.

Dimpfl, T. and Jank, S. (2011). Can internet search queries help to predict stock market volatility? Technical report, CFR working paper.

Dugas, A. F., Hsieh, Y.-H., Levin, S. R., Pines, J. M., Mareiniss, D. P., Mohareb, A., Gaydos, C. A., Perl, T. M., and Rothman, R. E. (2012). Google flu trends: Correlation with emergency department influenza rates and crowding metrics. *Clin Infect Dis.*, 54(4):463–469. PMID: 22230244.

Fama, E. F. (1965). The behavior of stock-market prices. *Journal of business*, pages 34–105.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work*. *The journal of Finance*, 25(2):383–417.

Fellbaum, C. (1998). *WordNet*. Wiley Online Library.

Graham, B. and McGowan, B. (2005). *The intelligent investor*. HarperCollins.

Jaffe, J., Randolph Westerfield, R., et al. (2005). *Corporate finance*. Tata McGraw-Hill Education.

Lütkepohl, H. (2007). *New introduction to multiple time series analysis*. Springer.

Lynch, P. S. (2000). *One up on Wall Street: how to use what you already know to make money in the market*. Simon and Schuster.

McTier, B. C., Tse, Y., and Wald, J. K. (2013). Do stock markets catch the flu? *Journal of Financial and Quantitative Analysis*, 48(03):979.

Ortiz, J. R., Zhou, H., Shay, D. K., Neuzil, K. M., Fowlkes, A. L., and Goss, C. H. (2011). Monitoring influenza activity in the united states: A comparison of traditional surveillance systems with google flu trends. *PLoS ONE*, 6(4):e18687.

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

Vozlyublennaia, N. (2014). Investor attention, index performance, and return predictability. *Journal of Banking & Finance*, 41:17–35.

Appendices

Appendix A

Table A.1: Sense ID from Word-Net and number of words selected

Sense ID	Number of Words	Number of random words selected
tops	85	8
act	11097	1109
animal	14780	1478
artifact	18743	1874
attribute	5707	570
body	3674	367
cognition	4882	488
communication	9309	930
event	1845	184
feeling	818	81
food	3762	376
group	4337	433
location	5261	526
motive	79	7
object	2383	238
person	21083	2108
phenomenon	1022	102
plant	18747	1874
possession	1633	163
process	1208	120
quantity	2241	224
linkdef	719	71
shape	565	56
state	5931	593
substance	4768	476
time	1833	183

Appendix B

Table A.2: Results from data handling Google trends data of random words from Word-Net

	USA	UK	AUS	CAN
Cases of Monthly data	3036	3458	3554	3611
Cases of non existing data	3493	4760	6422	5693
Cases with high number of zeros	1081	1748	1268	1252
Number of valid cases	6510	4154	2876	3564