Online Search Queries And Investor Sentiment: Financial Applications

Master's Thesis

for the attainment of

Master of Science in Economics and Business Administration

Finance and Strategic Management

at

Copenhagen Business School

2016

Author:

Lorenzo Tonelli

Supervisor:

Niklas Kohl

Department of Finance

Copenhagen Business School

Hand-in date:

1st June 2016

No. of characters (incl spaces):

127,079 (56 standard pages)

Abstract

This thesis examines the use of online search queries as a proxy for investor sentiment and it evaluates their ability to forecast (I) trading activity, (II) abnormal stock returns and (III) implied volatility. Prior research has highlighted that online search query data can be used to measure the attention of unsophisticated investors. Based on this insight, it is reasonable to expect search queries for traded companies to carry information which can predict financial market dynamics. The sample chosen consists of S&P500 constituents on a period ranging from 2007 to 2015. For each firm in this sample, search query data is obtained from Google in the form of Search Volume Index (SVI). Then, three studies are conducted in order to assess SVI's forecast capabilities. In study (I), the relation between SVI and trading activity is measured by computing a set of time-lagged cross correlation coefficients. In addition, a series of Granger-causality tests is conducted in order to ensure the robustness of the results. In study (II), SVI's capability to forecast abnormal returns is evaluated by simulating a series of long-short trading strategies based on SVI observations. Abnormal returns of each strategy are then computed by correcting for the most commonly recognized risk factors. In study (III), it is tested whether SVI can improve implied volatility forecasts: several implied volatility autoregressive models AR(p) are estimated in order to provide benchmark measurements; then, these models are augmented with SVI information. The forecasts produced by benchmarks and augmented models are compared and their accuracy is assessed with a series of indicators such as Mean Squared Prediction Error and Mean Absolute Percent Error, among others.

The three studies conducted indicate that (I) SVI anticipates and Granger-causes trading activity, (II) SVI incorporates information that translates in abnormal stock returns, but exploiting this phenomena is very difficult because financial markets quickly absorb such information and react accordingly. Lastly, (III) for given specifications of AR(p) models, SVI can improve implied volatility forecast both in-sample and out-of-sample.

Contents

Chapter 1

Introduction and problem statement

Over the last decades, the digital revolution has drastically transformed the way information is produced, accessed and processed. The widespread diffusion of digital devices such as personal computers and smartphones, coupled with the increasing availability of internet connections has radically modified what the preferred channels to share and gather information are. In a recent report, Perrin and Duggan (2015) provide evidence that from 2000 to 2015 the share of US adult citizens that make use of the internet rose from 52% to 84%. Internet adoption is even higher among individuals younger than 50, where the current usage rate has reached saturation levels that are above 93%.

In regards to information gathering, web search engines play a major role, as they allow users to quickly find the information they are looking for by scanning several sources and selecting the most relevant results among a plenitude of resources available. Due to their widespread adoption, most important search engine providers constantly receive massive amount of input from their users in the form of search queries submitted. Therefore, search queries constitute a valid proxy of internet population's attention. Given the high internet usage rate that characterizes some countries, it can be deducted that search queries capture the attention sentiment of a consistent part of their population

In order to improve the quality of the service provided, search results produced by web search engines make use of several pieces of information that users directly or indirectly transmit when they submit search queries: Google search results will vary considerably depending on the geographical location of the user, his search history, the browser he is using and several other parameters. After the search query is produced, the information transmitted by the user is not erased, but stored so that the search engine provider can improve its future services. All searches submitted to Google are made publicly accessible via Google Trends website[1] in an aggregated form, constituting what this paper will refer as the Search Volume Index (SVI). SVI represents the popularity of a selected word in relation to all other search queries submitted. Given that Google is the most popular search engine in the world, this index possesses characteristics that make it particularly appealing to researchers, because it can be interpreted as a timely measure of direct attention of a very significant part of the population. Clearly, there are several interesting applications for such index and

[1] https://www.google.com/trends/

its possible uses may comprise a wide array of research areas: several studies have highlighted SVI's capability to predict macro-economic indicators (Choi and Varian 2009; Dergiades, Milas, and Panagiotidis 2015) and to cast insights on different psychological traits of society (Guo, Zhang, and Zhai 2010). Researchers showed that SVI can anticipate the outbreaks of epidemics (Polgreen et al. 2008; Guzmán 2011), arguing that "harnessing the collective intelligence of millions of users, Google web search logs can provide one of the most timely, broad-reaching influenza monitoring systems available today"(Ginsberg et al. 2009).

Financial scientific literature has investigated SVI's capability to capture investor sentiment. Da et al. (2011), suggest that "search is a *revealed* attention measure: if you search for a stock on Google, you are undoubtedly paying attention to it. Therefore, aggregate search frequency in Google is a direct and unambiguous measure of attention". Based on this insight, Da, Engelberg, and Gao (2011) find evidence in their research that high SVI values can anticipate positive abnormal returns Furthermore, Bank, Larch, and Peter (2011) show that an increase in search activity seems to be followed by an increase in traded activity, stock returns and liquidity.

Inspired by the scientific literature aforementioned, this thesis investigates whether daily Google search volume can provide a timely measure for investor sentiment and examines its possible applications in the financial field. Three studies will be conducted in order to determine whether search queries can predict important financial indicators: (I) trading activity, (II) abnormal stock returns and (III) implied volatility. The firm sample analyzed is formed by all firms which composed the S&P500 index from January 2007 to November 2015. The search queries included in the analysis are divided in two types: company tickers and company "*topics*", where the notion "*topics*" refer to a refined query type that Google provides, which will be described in Chapter 3. Search query volumes are obtained from Google Trend website and for each firm it is built an index that measures abnormal SVI values (ASVI).

In Chapter 4, a positive correlation between ASVI and trading activity (proxied by abnormal stock turnover) is highlighted. The analysis is also expanded by conducting a series of Granger causality tests, which confirm the results. Moreover, these tests highlight that "SVI Granger-causes turnover" more often than "Turnover Granger-cause SVI". Lastly Granger-causality is also measured between SVI and realized volatility proxies, finding evidence that ASVI precedes increases in systemic and idiosyncratic risk.

In chapter 5, the capability of ASVI to predict future returns and abnormal returns is assessed. A preliminary analysis based on a rolling linear regression suggests that topic ASVI anticipates higher returns in the subsequent periods, but markets are very quick to incorporate such information. These findings are

confirmed in the second part of the study, where different trading strategies are simulated. The most unbiased and realistic trading strategy among the three proposed does not exhibit significant abnormal returns. It is concluded that ASVI carries information that can predict future returns, but exploiting such information seems unlikely.

In chapter 6, a benchmark AR(p) model is estimated in order to forecast implied volatility. The model is then augmented by the inclusion of ASVI variables, and forecast improvements are assessed with a series of indicators such as Mean Average Error, Mean Average Percent Error, Theil-inequality coefficient and Mincer-Zarnowitz R^2. Results indicate that the best performing in-sample models are capable to improve forecast accuracy out-of-sample, provided that an appropriate estimation window is chosen.

The analysis proposed differentiates from the existing literature in several ways: firstly, unlike for the vast majority of previous research, it is constructed a SVI index based on daily observations rather than weekly or monthly measures, resulting in a more timely indicator than the ones examined in previous studies. Furthermore, SVI information is also collected by "topics" instead of simple query terms, which should lead to increased data quality. Lastly, the use of SVI to forecast stock's implied volatility constitutes, to the best of my knowledge, an element of novelty among the existing literature.

Chapter 2

Literature review

Search query analysis tools have sparkled the interest of academics across multiple research areas. However, due to the recent nature of the service, existing literature is not over-abundant. The first part of this chapter provides an overview of research that has been conducted across non-financial academic fields, while the second part focuses on financial applications.

2-1 Online Search Behavior in general

Choi and Varian (2012) indicate Cooper et al. (2005) as the first study that makes use of search queries for forecasting purposes. In their research they use Yahoo! search query logs in order to estimate cancer incidence and mortality; the underlying assumption is that users are likely to use search engines in order to get information on their symptoms or in the attempt to gain better knowledge about diseases covered by media attention. Similarly, Polgreen et al. (2008) , Ginsberg et al. (2009) and Corley et al. (2009) use "flu related keywords" in order to predict influenza insurgency rates. With a similar approach, Prosper and Bangwayo-Skeetea (2015) use Google search queries to predict tourist influxes to the Caribbean and suggest that search queries should be used by policymakers and business practitioners in order to plan touristic service offer more efficiently. Similarly, Hand and Judge (2012) use query volumes to predict cinema admissions in the UK.

In terms of economic research, one of the earliest contributions can be attributed to Ettredge, Gerdes, and Karuga (2005), who find a positive and significant correlation between search queries and US unemployment rates. However, the data quality that was available at that time does not allow them to test search queries forecasting power. Askitas and Zimmermann (2009) build upon their research, using search queries to predict unemployment figures in Germany. Despite the low amount of data available for German search queries at that time, they provide evidence that search queries are reliable estimators of future unemployment rates. Similar results are also reported by additional studies who focus of different geographical areas, such as Italy (Francesco 2009; D'Amuri and Marcucci 2012) and Israel (Suhoy 2009). In addition, Choi and Varian (2009) suggest that Google Trends has the potential to help predicting unemployment benefit claims.

Other economic indicators have been shown to be linked to search queries: Guzmán (2011) suggests that Google trends data can serve as a predictor for inflation. Moreover, he shows that SVI significantly outperforms other most traditional indicators. McLaren and Shanbhogue (2011) show that Google insight data can be used to forecast economic indicators referred to labor and housing prices in the UK.

An interesting take on the use of Google trends come from Choi and Varian (2012) who claim that search queries can be used to *nowcast* economic indicators, i.e. observe current economic activity in real time, accessing information on events that already happened, but have not been made public by official reports yet. As an example, they claim that the search queries for an automobile maker firm can indicate the current level of its sales. Therefore, SVI can be used to predict the company's official sales report prior to its publication. Clearly, information of this kind would be precious for an investor, who could benefit from an indicator that is more timely than official statements.

2-2 Online Search Behavior and finance

Within the financial field, one of the earliest contributions can be attributed to Da et al. (2010). In their research, they use search queries referred to the most popular products of a set of companies, finding a positive correlation between SVI increase and firms' abnormal revenues. These results are shown to be consistent even after including a series of control variables such as firm size, market-to-book and historical returns; this leads the authors to argue that SVI is a relevant indicator that includes information which is not fully incorporated by the market. These findings will lead the authors to further expand their studies, presenting additional research on this topic.

Da et al. (2011) publish a study that, to date, constitutes one of the most extensive contributions to the SVI financial literature. They conduct an analysis on weekly search query volumes of Russel 3000 index components. The search queries used refer to the stock tickers of the companies in scope, after removing those tickers which are likely to be too noisy (e.g. "CAT, "BABY", "A" etc.). Da et al. (2011) show that SVI correlates with other measures of investors' attention such as news coverage and news events. In a vector autoregression (VAR) framework, they show that SVI anticipates such measures, which is consistent with the notion that investors may start to pay attention to a stock in anticipation of a news event. Moreover, the authors provide evidence that SVI captures the attention of individual/retail investors: they use SEC (Dash-5) monthly reports on retail order execution in order to distinguish between market centers that attract unsophisticated investors (e.g. Madoff) and market centers that attract sophisticated and institutional

investors (e.g. NYSE and Archipelago). They find evidence that market centers with more retail investors are more sensitive, in terms of trading activity, to an increase in SVI.

Next, the authors provide evidence that abnormal SVI values anticipate an increase in abnormal stock returns. Using a Fama-Macbeth regression approach and controlling for several alternative attention measures, the authors show that high SVI figures are followed by an outperformance of more than 30 basis points on a characteristic-adjusted basis during the subsequent two weeks. Lastly, the authors show that ASVI has significant predictive power for first-day IPO returns.

Further proof that high SVI anticipates positive abnormal stock returns is provided by Joseph, Babajide Wintoki, and Zhang (2011). Their analysis is builds on Barber et al. (2008) and Schmeling (2007), who suggest that investor sentiment forecasts stock returns. They construct five portfolios which are sorted weekly bades on stock's SVI on the previous week. Thus, they derive a long-short strategy which goes long on high SVI stocks and shorts low SVI stocks. Every week, the long-short portfolio is rebalanced according to the new SVI figures. After controlling for appropriate risk factors, this portfolio exhibits abnormal returns of about 7% annually.

The aforementioned results are consistent with a similar study conducted on the German Market by Bank, Larch, and Peter (2011). In addition, the authors find a positive correlation between SVI and other measures of trading activity such as stock turnover and illiquidity ratio.

SVIs has also been proven to be capable to predict stock volatility, where the most notable contribution can be attributed Dimpfl and Jank (2012). The authors find evidence of a strong co-movement of SVI index with Dow Jones realized volatility. In particular, they show that including SVI information to autoregressive forecasting models of realized volatility significantly improve the forecast. This effect is particularly strong during high volatility phases.

Chapter 3

Data

This chapter discusses the data used, its limitation and the choices made in the sample construction phase. The Chapter is structured as follows: sections (3-1) to (3-4) present some consideration on SVI data availability, limitations and interpretations. Section (3-5) describes the approaches followed in order to download financial figures and SVI data. Lastly, section (3-6) describes how some variables were manipulated in order to obtain indicators that will be used for the analysis.

3-1 Search engine data availability

Although Google Trends is the most popular and widely used search analysis tool, other providers offer (or have been offering) services of similar nature. Furthermore, the quality of publicly available search query data available has constantly improved over time, making today's data more accurate and accessible. This section discusses the main alternative services to Google Trends and presents a brief overview of the improvements that have been implemented to Google trends ever since its inception.

In recent years, all major Search engine websites (Google, Yahoo and Bing) have provided free public tools that allow to analyze the search queries submitted. The primary purpose of such tools was to help webmasters and advertisers to better direct their products to the public: the popularity of several keywords could be confronted and measured, leading to a more efficient SEO[2] writing.

Search engine *Bing* offers a *Keyword Research Tool,* while on November 2010 *Yahoo!* Introduced a service called *Yahoo! Clues* that allowed to gain insight regarding the search volumes. Moreover, the reports included demographic information on the user, such as age, gender, location etc.. *Yahoo!* searches have been analyzed by Yi, Maghoul, and Pedersen (2008), Ginsberg et al. (2009), Ricardo Baeza-yates (2007) and (Rose and Levinson 2004). Moreover, (Bordino et al. 2012) discuss the correlation between *Yahoo!* Queries and stock traded volume. Unfortunately, *Yahoo! Clues* was discontinued in April 2013 as a result of a company restructuring aimed to sharpen the strategic focus of the organization.

[2] Search Engine Optimization: the process of maximizing the number of visitors to a particular website by ensuring that the site appears high on the list of results returned by a search engine.

On May 2006 *Google* introduced *Google Trends*, giving users the possibility to analyze the volume of the queries submitted to the search engine. Given the massive market share of *Google* over its competitors, it is clear why *Google Trends* immediately appeared particularly appealing among the scientific community: it was now possible to obtain insights generated by a sample of about 70% of Internet population[3]. During its first stages *Google Trends* posed several limitations to its users: on the go live date it was possible to analyze data starting from 2004 but new data was not added real time. Sporadic updates added blocks of data in the upcoming months, but there was not a clear update schedule set. Moreover, it was neither possible to download data nor to filter the report by user location, making data analysis difficult and impractical.

On August 2008, *Google* announced the introduction of a new tool called *Insights for Search*[4] aimed to help webmasters to better understand search behavior. *Insights for Search* allowed to setup geographical filters and to analyze the correlation between queries.

On September 2012 *Google Insights for Search* was merged with *Google Trends* resulting in a new website that combined the features of both services: it was now possible to select one or more query words and setup geographical filters, time scope filters and more.

3-2 Google Trends Search Volume Index (SVI) interpretation

Google Trends does not allow to obtain the absolute number of searches for a given keyword (Dergiades, Milas and Panagiotidis 2015). Instead, the website returns scaled numbers, that are commonly referred as *Search Volume Index* (SVI).

Google provides the following explanation on how to interpret the SVI[5]:

> *"The numbers that appear show total searches for a term relative to the total number of searches done on Google over time. A line trending downward means that a search term's relative popularity is decreasing. But that doesn't necessarily mean the total number of searches for that term is decreasing. It just means its popularity is decreasing compared to other searches. […] Numbers represent search interest relative to the highest point on the chart. If at most 10% of searches for the given region and time frame were for "pizza," we'd consider this 100. This doesn't convey absolute search volume."*

[3] http://www.comscore.com/Insights/Rankings/comScore-Releases-April-2014-US-Search-Engine-Rankings
[4] http://adwords.blogspot.dk/2008/08/announcing-google-insights-for-search.html
[5] https://support.google.com/trends/answer/4355164?hl=en&rd=1

More formally, the procedure the website uses in order to obtain the SVI can be described as follows: let ASV_i^t denote the *Absolute Search Volume* of keyword *i* at time *t* (being *t* a month, week, day. etc., depending on the report extracted) and let ASV_{Total}^t denote the Absolute Search Volume of all search queries submitted to the search engine on the same time interval.

The Normalized Search Volume (NSV) of keyword *i* can then be written as follows:

$$NSV_i^t = \frac{ASV_i^t}{ASV_{Total}^t} \tag{3.1}$$

NSV_i^t is then scaled in order to obtain SVI_i^t, that is the only measure Google publicly discloses:

$$SVI_i^t = \frac{NSV_i^t}{\max(NSV_i^{t_l}, \dots, NSV_i^{t_u})} * 100 \tag{3.2}$$

With t_l and t_u denoting respectively the lower and upper bound of the time interval included in the report. As it can be seen, it will always hold that $SVI_i^t \leq 100$.

It is also possible to request a report that includes more than one keyword, in this case the scaling factor (denominator of equation (3.2)) will be the highest NSV_i^t across all keywords and time intervals included. However, multiple keyword reports will not be used in this paper. Note that, among all terms of equations (3.1) and (3.2), SVI is the only variable that is disclosed to Google Trends users.

3-3 SVI extraction limitations

In the previous paragraph, it has been mentioned that SVI's observation frequency *t* can be represented by months, weeks, days, hours or even shorter intervals. However, it must be noted that the user cannot explicitly select the time frequency of the observations in the report. In order to obtain daily observations, two conditions must be met:

1) $t_u - t_l \leq 93\ days$ (3.3)

Any report with a timespan greater than 3 months will always generate weekly or monthly observations.

2) $ASV_i \geq L_d$ (3.4)

The Absolute Search Volume of keyword *i* throughout the selected time interval must be greater than an arbitrary threshold value L_d set by Google. In addition to L_d , there are other threshold values that determine whether the report can be produced with weekly or monthly observation frequency. If we denote them by L_w and L_m respectively, we can write that $L_d > L_w > L_m$. Therefore, keywords with a high ASV_i will be reported with daily observations, while less commonly used keywords will be reported with weekly or monthly observations

It is important to note that both terms L_d and ASV_i in condition (3.4) are not directly observable by the user. Only submitting the query and observing the output can reveal whether the condition is satisfied.

As of today, Google has not release an API[6] for Google Trends. This limits the accessibility of the platform, as users have to rely on manual extraction or unofficial web-crawling solutions. As further sections will explain, for this research a web-crawling solution was developed in order to obtain the reports in a structured and automated manner

3-4 An empirical example

In an attempt to help the reader to better understand the data available on google trends, an empirical example is presented: picture 3-1 shows the results of a query referred to bank of Bank of America's stock ticker ("BAC"). The geographical scope has been restricted to United States, in order to limit the noise deriving from other unrelated terms that could be described by acronym "BAC".

[6] Application programming interface: a set of functions and procedures that allow the creation of applications which access the features or data of a service.

The bottom right panel shows the most related queries to our ticker, i.e. the most common queries submitted by users who also searched "BAC": the fact that "bac stock" is among them is encouraging. However, we can also see that there is a certain level of noise in our report: a quick investigation reveals that related search "the bac" refers to the *Boston Architectural College*, while "bac calculator" probably refers to "Blood Alcohol content calculator". The existence of noisy terms like the ones just presented is very common, and hinders SVI's capability to capture financial sentiment.

The SVI graph clearly shows an interest spike on the first half of year 2009. This probably refers to the increased investors' attention during the US recession. The peak in April 2009 (marked with (A)) coincides with the Bank's earnings announcement release. A second peak occurs in the second half of August 2011 (B). The cause of this peak is uncertain; however, the existence several articles released on that period referred to Warren Buffet decision to invest $5 billion in Bank of America[7] may constitute a possible explanation.

[7] http://dealbook.nytimes.com/2011/08/25/buffett-to-invest-5-billion-in-bank-of-america/
http://investor.bankofamerica.com/phoenix.zhtml?c=71595&p=irol-newsArticle&ID=1600359#fbid=Uc6BC7HQWcO

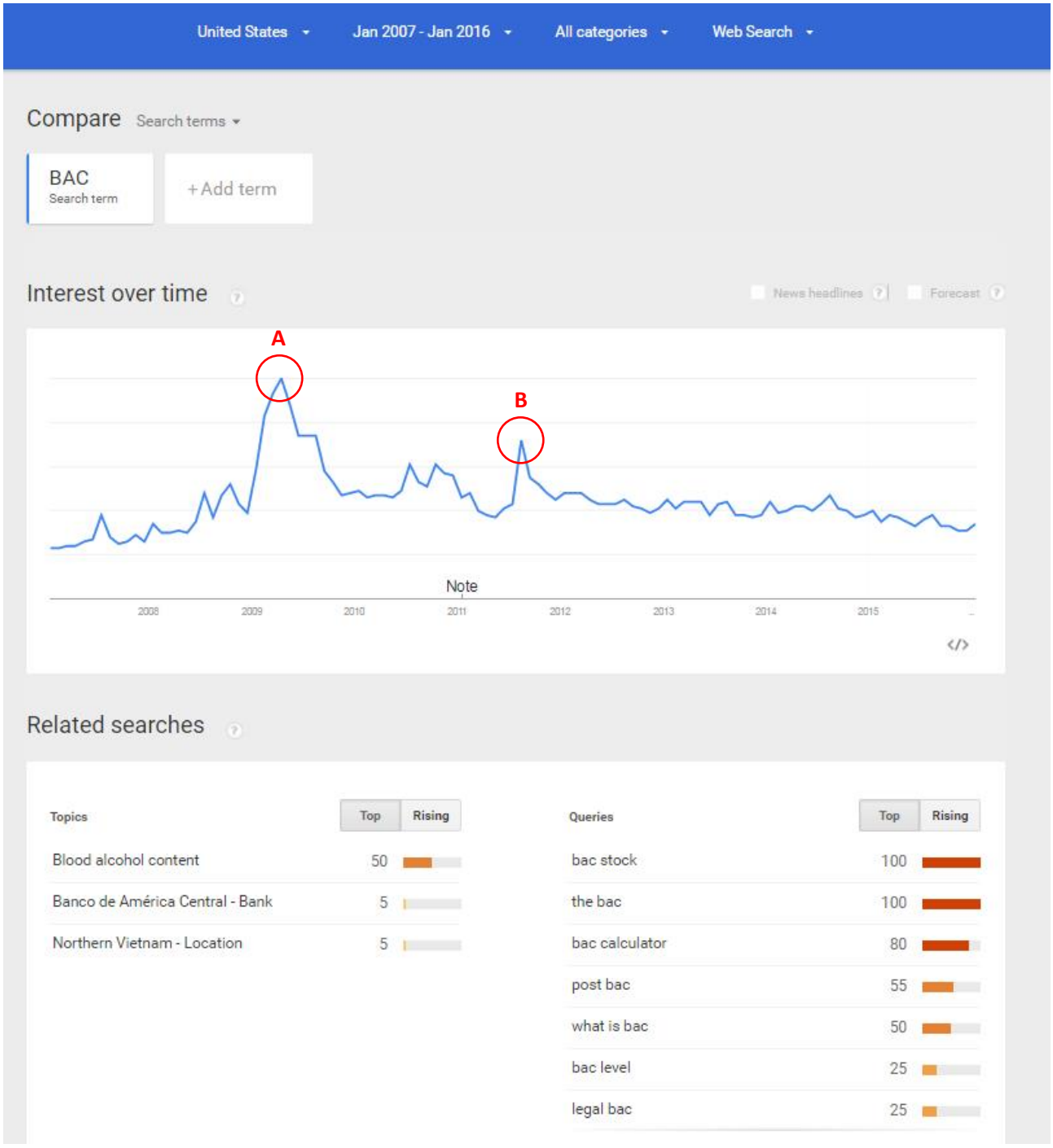


Figure 3-1: Google Trends report for string "BAC". Query is filtered by geographical area (United States) and time period (Jan 07 – Jan 16), Report includes SVI measurements over time, as well as related searches and topics

3-5 Sample Construction

3-5-1 Financial data

The primary financial variables used are listed in table 3-1 below. The rest of this section describes the methodology used for dataset construction.

Table 3-1: list of financial variables downloaded

	Variable	Notation	Source
-	Ex-dividend returns	r_t	CRSP
-	Opening prices	P_t^{Open}	CRSP
-	Outstanding Shares	OS_t	CRSP
-	Traded Shares	T_t	CRSP
-	Implied volatility	IV_t	Bloomberg
-	Cumulative factor to adjust prices	cfacpr	CRSP
-	Fama-French risk factors + UMD	MKT,HML, SMB, UMD	FF website[8]

The study sample consists of all firms which composed the S&P 500 index between January 2007 and November 2015. In order to avoid survivorship bias, the sample includes all companies which have been part of the index at any time during the selected time period. When a company ceases to be a S&P 500 constituent but its stocks are still being traded, it is not removed from the sample, i.e. it is assumed that investors are not affected by this information and will keep trading it.

The components' list was obtained from Compustat, while the daily observations of relevant variables were downloaded from CRSP. CRSP\Compustat merged database[9] was used in order to match the primary identifiers of the two datasets (GVKey for Compustat and LPermNO for CRSP), this merging procedure is referred as "best practice" according to Wharton's WRDS user manual[10].

After merging the datasets, the sample size has increased because CRSP assigns a new LPermNO for every new stock issue. Therefore, it can be that a single company (characterized by one single LPermCO) belongs

[8] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html
[9] http://www.crsp.com/products/research-products/crspcompustat-merged-database
[10] https://wrds-web.wharton.upenn.edu/wrds/support/Data/_001Manuals%20and%20Overviews/_002CRSP/ccm-overview.cfm

to several LPermNOs. As it will be shown later, the SVI of this study consist of stock tickers and company names; therefore, it is reasonable to compact the database using these terms as identifiers. It should be noted that tickers can be recycled and therefore they do not necessarily constitute a unique identifier; this does not constitute an issue when associating the time series of ticker SVI, as it is reasonable to assume that investors will google the ticker of the company that held that ticker code at that moment of time. However, when the ticker is based on the company name, the companies obviously need to be kept distinct and matched with an appropriate search query term. The initial sample size is composed by 699 firms; it is important to ensure that the number of cross sectional units is high, because as section 3-3 will explain, several technical restrictions to SVI availability will greatly reduce the size of the usable sample.

Being very popular stocks, data quality of all the time series mentioned above is expected to be rather high. However, some outlier observations were observed in stock opening prices. A sanity check conducted using Yahoo! Finance highlighted a few discrepancies, that were corrected manually. (cf. Appendix A1 for a representation of the time series outliers). After the series is corrected, opening prices are used to compute open-to-open returns (r^{open}) that will be used in Chapter 5.

In addition to CRSP data, Bloomberg database was used in order to retrieve daily observation of stock's implied volatility[11]. For US stocks, Bloomberg provides the weighted average of the implied volatilities of the closest out-of-the-money call options and the closest out-of-the-money put options.

3-5-2 Google Trends query terms considerations

Choosing an appropriate query word is essential in order to correctly capture the investors' sentiment. Under this perspective, existing scientific literature suggests two main approaches: searching the company by names or by company tickers (Da et al. 2011). Both approaches come with advantages and disadvantages and previous literature is not unanimous regarding which one performs better: Da et al. (2011) argue that using company tickers provides a better estimate of abnormal return, while Vlastakis and Markellos (2012) opt for company names, Latoeiro, Ramos, and Veiga (2013) suggests a mixed approach. The rest of this section we discuss some considerations on the different approaches.

[11] The corresponding functions are HIST_CALL_IMP_VOL and HIST_PUT_IMP_VOL, which for US stocks return the same value (average)

Approach 1: Search by company name

Using company names generally returns higher amounts of data, returning in more valid reports that satisfy equation (3.4). Nevertheless, data availability depends on the specific company: firms with simple names that are generally written with no variants (e.g. "Amazon") will generally return valid results. However, some firm names can be written in several ways[12] and it is difficult to predict users' behavior. In this case, the total search query volume will be distributed among several similar keywords, decreasing the chance of satisfying equation (3.4).

Moreover, it may be argued that searching by company name generally returns noisy results, because this approach captures the interest of all users, not only those that are interested in financial information. This is particularly true for retail companies and online stores; whose name is likely to be googled by their customers. Picture 3-2 provides an example of this issue, showing the discrepancy between the web interest for "*Amazon"* (characterized by peaks on the Christmas holiday periods) and its ticker, which has a much lower volume and follows a different pattern.

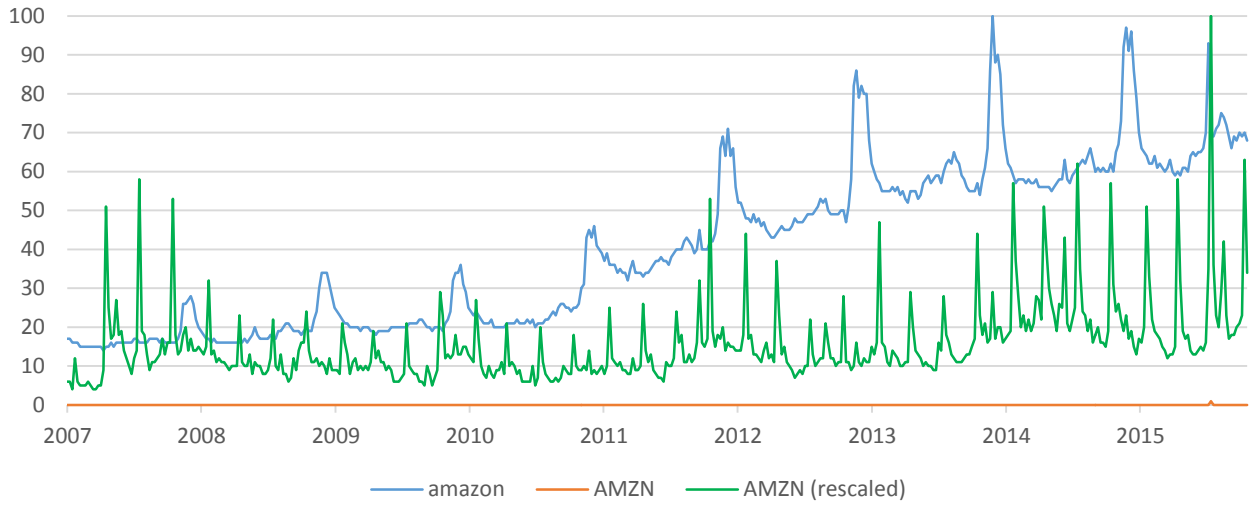


Figure 3-2: Adjusted Search Volume Indexes for string "Amazon" and ticker "AMZN" over time. The SVI of the latter terms is much smaller than the string term and therefore it has been rescaled to ease the comparison. "Amazon" string greatly increases as the Christmas holiday season approaches. String "AMZN" follows a very different pattern. This picture shows that the two queries capture different information.

[12] As an example, "Bristol-Myers Squibb Pharmaceuticals" may be also written "Bristol Myers Squibb", "Bristol Myers pharma", "BM Pharmaceuticals" etc.

An argument in favor of using company names instead of tickers is that company names are less prone to misinterpretation: search term "Pfizer" will almost surely refer to the pharmaceutical company, while Pfizer's ticker "PFE" may be the acronym of many unrelated terms.

Approach 2: Search by company ticker

The main argument in favor of using stock tickers as query terms is that they may be able to better catch the interest towards the *financial* characteristics of a company. However, tickers can also be very noisy: short combinations of letters can represent a multitude of acronyms and short words. The most common approach is to manually remove tickers that are clearly noisy (Da, Engelberg, and Gao 2011), but this is not always easy, as the researcher may not be aware of some technical acronyms that constitute noise. Moreover, particular care has to be exerted in order to avoid bias distortions.

Approach 3: Search by company topic

Google offers a feature that allows users to search *topics*, i.e. to aggregate searches referred to the same topic or entity based on semantic criteria. Google official blog explains this feature as follows[13]:

> *"When you measure interest in a search topic (Tokyo - Capital of Japan) our algorithms count many different search queries that may relate to the same topic (東京, Токио, Tokyyo, Tokkyo, Japan Capital, etc). When you measure interest in a search query (Toyko - Search term), our systems will count only searches including that string of text ("Tokyo")"*

It is clear that this approach could represent a more refined version of Approach 1, because it filters out noisy and ambiguous terms. In addition, because several keywords referred to the same topic are consolidated, the odds of satisfying condition (3.4) increase considerably, resulting in a much higher number or valid reports. Google states that the topic classification of searches is subject to continuous improvement: for the sake of this study, it is important to ensure that these improvements are not applied retroactively, as this may result in *look-ahead bias.* To the best of my knowledge, the only retroactive change has occurred on 1/1/2011, when the geographical assignment of searches within the USE was improved, and a public statement was

[13] https://search.googleblog.com/2013/12/an-easier-way-to-explore-topics-and.html

introduced by google to notify users and researchers. Given that this paper is primarily concerned with the information content of SVI and not its geographical origins, this does not seem to constitute a relevant issue. Moreover, Choi and Varian (2012) describe Google's categorization as a reliable tool for econometric analysis.

Approach chosen in this study

Based on the arguments presented above, it is clear that each approach comes with advantages and disadvantages. However, approach (3) based on *topic SVI* seems to dominate the approach (1) based on *name SVI*. In the earlier stages of this research, SVI reports for all three approaches were downloaded, but it was soon clear that "name SVI" approach was not producing enough valid reports. Therefore, only *ticker SVI* and *topic SVI* have been used.

It is important to note that the two different SVI types will never be mixed in the study. Instead, they will be handled as two distinct indicators. The reason for this approach is twofold: firstly, it allows to compare the two measures and to establish which one performs better. Secondly, it eliminates the need to make arbitrary decisions pertaining which indicator to use for each firm.

3-5-3 SVI extraction and limitations

Section 3-5-1 described how a sample of 699 firms based on the S&P500 index components was determined. For each firm, ticker SVI and topic SVI have been downloaded. The rest of this section describes the procedure followed in order construct the SVI dataset.

A web-crawling software was used in order to download reports covering a time interval of 3 months each, with an overlap of one month over each other. This procedure is essential in order obtain reports with daily observations (cf. condition (3.3)). Because the sample ranges from January 2007 to November 2015, for each search term 53 reports are downloaded.

Since the firm sample is composed by US companies, reports were filtered in order to include only queries generated in the United States. This is consistent with the intuition that investors prefer to invest in local firms, because they have better knowledge of the local market (Ivković and Weisbenner 2016). This intuition is also confirmed by Da et al. (2011), who argue that US queries yield to better information regarding the US stock market.

In order to download topic data, every company is searched on Google Trends, and the topic that better associates with the firm in scope (i.e. the company name) is selected. Occasionally, Google Trends does not return any topic that could match the company name, in this case the firm is excluded from the sample. Moreover, some topics do not return valid SVI figures due to low search volumes, making them not usable. As a result of this process, the *topic SVI* firm sample is reduced to 411 companies (21783 reports).

When downloading ticker data, a series of filters are applied in order to eliminate noisy tickers: tickers composed by a single letter (13 tickers) are removed from the sample, as well as tickers that represent an unrelated word (e.g. CAT, ALL etc., 39 tickers). Among the remaining tickers, many of them do not contain data because the query parameters do not satisfy condition (3.4) during some intervals the sample or for the whole sample period. After removing the series with missing values the *ticker SVI* firm sample is composed by 122 companies (6466 reports).

3-5-4 Merging Google Trends reports

Previous section described the procedure used to select and download SVI reports. It has been mentioned that for each firm 53 reports are obtained in order to cover the full sample period. This section describes how the reports are merged in order to obtain a single time SVI series per firm.

Each report is composed by approximately 90 daily observations SVI_t. For each report it is then computed the percent change $C_t = (SVI_t/SVI_{t-1}) - 1$. As mentioned earlier, each report overlaps with the next one for the duration of one month. Therefore, during the overlap periods it is possible to compute two values of C_t, one for each report: one would expect these values to be identical, because they refer to the same Absolute Search Volume ASV_t and they result in different SVI_t only because of the different normalization term of the series. However, I find that the corresponding C_t values of overlapping periods are sometimes different, and they occasionally even have different signs. One possible explanation could lie in the fact that, in order to increase response speed, when Google Trends generates a report, it selects a random subsample of the actual historical search data and it calculates the SVI based on that subsample. Therefore, extracting the same report multiple times does not always lead to identical results. In order to measure the magnitude of this distortion, a random subsample of tickers and topics are selected and several copies of the same reports are downloaded multiple times. The correlation between the reports never fell below 0.98, indicating that the distortion produced by Google's sampling procedure is very small and it should not affect our study's estimates. Da et al. (2011) run a similar sanity check on their data and obtain similar results.

In order to produce a single SVI series for each cross section, each series is initiated with an arbitrary value of 1 and for each period *t* the merged SVI (SVI_t^*) is computed based on the percent change C_t More formally:

$$SVI_t^* = \begin{cases} 1, & t = 1 \\ SVI_{t-1}^* \cdot C_t, & t > 1 \end{cases} \tag{3.5}$$

Lastly, each series is normalized for an easier interpretation, so that each observation will lie between 0 and 100:

$$SVI_t^{**} = \frac{SVI_t^*}{max(SVI_1^*, \dots, SVI_T^*)} * 100 \tag{3.6}$$

3-6 Variable construction

Variables described so far need to be manipulated in order to obtain useful variables for the study. Namely, it is necessary to define and compute abnormal values of SVI, traded volume and stock returns. This section describes the approach adopted.

3-6-1 Abnormal SVI

D et al. (2011), suggest to use the following variable in order to measure abnormal SVI values:

$$ASVI_t = \log(SVI_t^*) - \log[Median(SVI_{t-1}^*, \dots, SVI_{t-8}^*)] \tag{3.7}$$

Where, $\log(SVI_t^*)$ is the logarithm of SVI* on week *t* and $\log[Median(SVI_{t-1}^*, \dots, SVI_{t-8}^*)]$ is the logarithm of the median SVI* during the prior 8 weeks. They explain that "*the median of the time window captures the normal level of attention in a way that is robust to recent jumps*" and they argue that another advantage of this measure is that it removes time trends. The dataset of this study is composed by daily observations, therefore time window *k* should be adjusted. Thus, equation 3.7 can be rewritten as:

$$ASVI_t^k = \log(SVI_t^*) - \log[Median(SVI_{t-1}^*, \dots, SVI_{t-k}^*)] \quad (3.8)$$

$ASVI_t^k$ is computed for *k = 5,10,20,40,60.*

3-6-2 Abnormal stock turnover

Stock turnover T_t^i on day *t* is defined as:

$$T_t^i = \frac{TV_t^i}{OS_t^i} \quad (3.9)$$

Were TV_t^i is the number of shares of firm i traded and OS_t^i is the number of shares outstanding. Wang (1994) and Lo and Wang (2000) provide a theoretical justification for using turnover instead of other raw volume metrics. The main advantage of indicator T_t^i versus the simple traded volume is that T_t^i is adjusted for stock splits. Abnormal turnover AT_t^k is derived using the same approach suggested for abnormal SVI:

$$AT_t^k = \log(T_t) - \log[Median(T_{t-1}, \dots, T_{t-k})] \quad (3.10)$$

AT_t^k is computed for *k = 5,10,20,40,60*

3-6-3 Abnormal stock returns

Abnormal returns are defined as the difference between realized and expected returns. More formally, the abnormal returns of stock *i* at time *t* can be written as:

$$ar_{i,t} = r_{i,t} - E(r_{i,t}|X_t) \quad (3.11)$$

Where $r_{i,t}$ denotes the realized returns and X_t is the set of information available up to *t*, which is used to compute the expected returns.

Computing abnormal returns poses to the researcher the choice pertaining which model to use. According to arbitrage pricing theory, expected returns of a financial asset can be modeled as a linear function of relevant risk factors. Perhaps the most common approach is to compute abnormal returns according to Fama-French 3-Factor Model (FF3)(cit. x). The first term of the model (MKT-RF) is common to the CAPM theory and represents the excess return on the market. It is calculated as the value-weight return on all NYSE, AMEX, and NASDAQ stocks (from CRSP) minus the one-month Treasury bill rate (from Ibbotson Associates). The second term (SMB, Small Minus Big) is the average return on the three small stock portfolios minus the average return on the three large stock portfolios. The third factor (HML, High Minus Low) is the average return on two high book-to-market portfolio minus the average return on the two growth portfolios (i.e. low book-to-market). $r_{f,t}$ denotes the risk free rate.

Therefore, expected returns can be modelled as follows:

$$E(r_{i,t}|X_t) = r_{f,t} + \beta_{1,i}(r_{m,t} - r_{f,t}) + \beta_{2,i}HML_t + \beta_{3,i}SMB_t + \epsilon_{i,t} \quad (3.12)$$

The weights ($\beta_1, \beta_2, \beta_3$) are used to compute the abnormal returns in the following period *t+1* and they are calculated for each period *t* using a rolling window of one year (250 trading days). Therefore, it can be written that:

$$ar_{i,t} = r_{i,t} - (r_{f,t} + \beta_{1,i}(r_{m,t} - r_{f,t}) + \beta_{2,i}HML_t + \beta_{3,i}SMB_t) \quad (3.13)$$

Where $ar_{i,t}$ denotes the abnormal returns of stock *i* at time *t*.

3-7 Sample description

Section 3-5 described the procedures used in order to download and merge the data, while section 3-6 described how the data was manipulated in order to produce the time series that will be used in this study. This section presents descriptive statistics of the data presented so far.

Figure 3-3 shows the average value of normalized and abnormal SVIs in the two variants downloaded: *ticker* and *topic*. Both series exhibit negative peaks at the end of the year during the Christmas holiday. Clearly, the collective attention during the holiday period changes, and company information is less sought than usual. Other negative peaks occur during US bank holidays. Abnormal SVI variables seem to remove the trends that characterize portions of the SVI series. However, the distribution plots suggest that the series still deviate from the condition of normality.

Average topic SVI shows a jump between year 2010 and 2011. After this jump occurs, the average SVI seems to maintain a higher value than previous years. This jump occurs on the day google adjusted the geographical assignment of search queries (cf. Section 3-5-2). This adjustment is the most likely cause of the discontinuity.

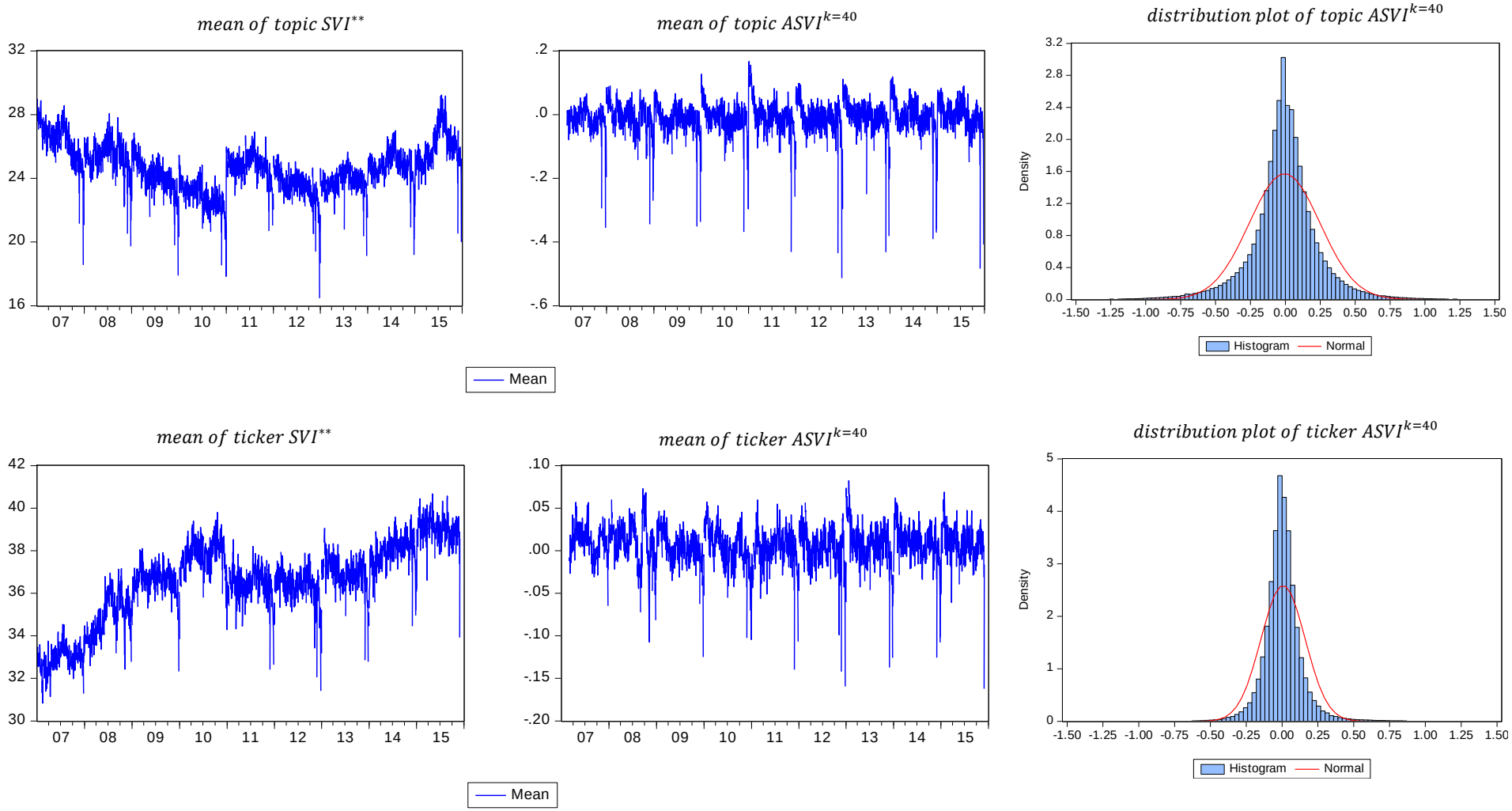
Table 3-2 reports descriptive statistics for selected time series. Abnormal time series are calculated according to the procedures described in section 3-6. Abnormal SVI are displayed for *k=40*, but different estimation windows exhibit similar characteristics.

Except for the normalized ticker SVI, the kurtosis of all series are higher than the value expected under normality assumptions (3). This suggests "fat tailed" distributions, which is a fairly common characteristic of financial time series. All series are characterized by positive skewness, which indicates that their distributions are asymmetrical: the tail on the right side is longer or fatter than the tail on the left side. The JB statistics strongly rejects the hypothesis of normality for all series.

Table 3-2: descriptive statistics of selected time series. The table reports figures for (left to right): normalized topic SVI (SVI^{**}), abnormal topic SVI ($ASVI^{k=40}$), normalized ticker SVI, abnormal ticker SVI, ex-dividend returns (r), open-to-open stock returns (r^{open}), abnormal stock returns (ar), stock turnover (T), abnormal turnover rate ($AT^{k=40}$), implied volatility (IV) and its natural logarithm.

	Topic		Ticker		Returns			Volume		Implied volatility	
	SVI^{**}	$ASVI^{k=40}$	SVI^{**}	$ASVI^{k=40}$	r	r^{open}	ar	T	$AT^{k=40}$	IV	$\log(IV)$
Mean	24.69	-0.004945	36.55	0.007175	0.000507	0.000494	-4.08E-05	11.48	0.023683	33.06	3.389
Median	20.03	-0.000224	35.01	0.000000	0.000389	0.000464	-0.000282	8.237	-0.007203	28.46	3.348
Maximum	100.0	4.605159	100.0	3.506558	2.750000	5.277778	2.583178	1107	11.92513	1064	6.969
Minimum	0.022	-2.370244	0.398	-2.352309	-0.839506	-0.740169	-0.830661	0.000	-5.993961	1.083	0.079
Std. Dev.	19.10	0.253507	20.94	0.154121	0.025143	0.026462	0.019249	13.22	0.448006	18.42	0.443
Skewness	0.941	0.207262	0.271	2.012181	2.161262	10.99598	3.912325	12.82	0.668757	3.355	0.615
Kurtosis	3.300	10.82410	2.297	35.05400	181.9865	1776.899	429.7097	480.0	7.993183	32.86	3.768
Jarque-Bera	135727.6	2251683	10894.88	14184651	1.47E+09	1.49E+11	7.30E+09	1.04E+10	1196566.	42615201	95939
Probability	0	0	0	0	0	0	0	0	0	0	0

Figure 3-3: cross sectional mean values over time and distribution plot of topic SVI, ticker SVI, topic ASVI and ticker ASVI (w=40)



Chapter 4

SVI and trading activity

The purpose of this section is to assess whether the SVI carries information that can be used to anticipate and forecast market activity; the reasoning behind this intuition is that investors do presumably need to gather information before undertaking investment decisions and SVI should be capable to capture this increase of information demand. Since ASVI is a measure of global attention, its level should increase when the demand for information on the firm is high, and such increase should be followed by increased trading activity: investors trade after they have consumed the information. In order to test this hypothesis, the following approach is adopted: firstly, pairs of time-series {Q, V} are selected, where V denotes a variable used as a proxy for traded volume and Q indicates a proxy of search queries. Subsequently, time lagged correlation coefficients between the pairs' members are computed. This process is repeated with several pairs {Q, V} in order to ensure the robustness of the results. As a second approach, a series of Granger-Causality tests is run on pairs {Q, V} in the attempt to assess whether Q carries predictive information over V and vice-versa. This analysis is further expanded by including proxies of Idiosyncratic and Systematic risk in the series pairs {Q, V}. The rest of the chapter presents the methodology used, followed by the analysis of the results.

4-1 Methodology

4-1-1 Lagged Cross Correlations

Correlation is measured by computing the time-lagged Pearson cross correlation coefficient:

$$r(\delta) = \frac{\sum_{t=1}^{n}(Q_t - \bar{Q})(V_{t+\delta} - \bar{V})}{\sqrt{\sum_{t=1}^{n}(Q_t - \bar{Q})^2}\sqrt{\sum_{t=1}^{n}(V_{t+\delta} - \bar{V})^2}} \tag{4.1}$$

Where Q and V denote the search query volume and a stock trading volume series respectively, and δ indicates the lag order used. Measures of trading activity have been described in section 3-6-2: variable used will consist of log stock turnover $\log(T_t^i)$ and several specifications of the abnormal stock turnover AT_t^k calculated using different time windows *k = 5,10,20,40,60*. Similarly, web query volumes are measured by $\log(SVI_t^*)$ and $ASVI_t$.

If the SVI is able to predict trading activity, not only Q should be correlated with T, but positive lag orders should also exhibit higher correlation coefficients. However, it is well known that a significant part of stock trading transactions is carried out by professional investors, which can generally rely on more specialized and efficient information channels. For this reason, SVI is expected to measure only the interest of unsophisticated investors. These investors are generally slower in acquiring and processing information, and may react after the market. This consideration limits the expectations regarding the predictive power of SVI, which may only capture the sentiment of "slow" traders who act aftermarket shocks have already taken place.

4-1-2 Granger Causality Test

Granger causality test is a statistical hypothesis test, widely used in time-series analysis in order to determine whether one series is useful in forecasting another. Granger-causality relationship lies on two fundamental principles:

1) The cause happens prior to its effects
2) The causal series contains unique information about the series being caused

A variable X is said to *Granger-cause* another variable *Y* if the predictions of *Y* based on the lagged values of *Y* and *X* are more accurate than the predictions of *Y* based solely on the lagged values of *Y*. It is important to note that the statement "*X Granger-causes Y*" does not imply that *Y* is the result of *X*: Granger-causality test does not assess causality in the more common use of term, the test only measures precedence and information content.

In order to conduct a Granger causality test of lag order l, observations of Y are regressed on their lagged values; i.e. the following autoregression is estimated:

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_l y_{t-l} + \varepsilon_t \tag{4.2}$$

Then, the autoregression is augmented by including the lagged values of the second variable *X* of which we want to test the explanatory power: the following bivariate equation is estimated:

$$y_t = \alpha_1 y_{t-1} + \cdots + \alpha_l y_{t-l} + \beta_1 x_{t-1} + \cdots + \beta_l x_{t-l} + \varepsilon_t \quad (4.3)$$

Then, it is computed the F-statistics for the joint hypothesis:

$$\beta_1 = \beta_2 = \cdots = \beta_l = 0 \quad (4.4)$$

Where the null hypothesis H_0 is that *X does not Granger-cause Y.* If the p-value of the F-statistics lies below the significance level chosen, null hypothesis is rejected, and it is concluded that *X* Granger-causes *Y*, i.e. the coefficients of variable *X* in equation (4.3) are statistically different from zero.

The same procedure is then repeated in order to test whether Granger causality exists in the opposite direction: similarly to equations (4.2) and (4.3), the following equations are estimated:

$$x_t = \alpha_1 x_{t-1} + \alpha_1 x_{t-2} + \cdots + \alpha_l x_{t-l} + \varepsilon_t \quad (4.5)$$

$$x_t = \alpha_1 x_{t-1} + \cdots + x_l y_{t-l} + \beta_1 y_{t-1} + \cdots + \beta_l y_{t-l} + \varepsilon_t \quad (4.6)$$

And the test for the joint hypothesis is:

$$\lambda_1 = \lambda_2 = \cdots = \lambda_l = 0 \quad (4.7)$$

Where this time the null hypothesis H_0 is that *y* does not Granger-cause *x.*

In order to test whether abnormal SVI values has predictive power over abnormal turnover, Granger causality test is run on each pair $\{Q_i, V_i\}$, where Q and V denote the search query and turnover series

respectively for each stock *i*. If SVI Granger-causes turnover, one would expect H_0: *SVI does not Granger-cause turnover* to be rejected more often than H_0: *turnover does not Granger-cause SVI*

Lastly, It is tested SVI's ability to forecast systematic and idiosyncratic risk. The proxy indicators used for this purpose are $|r_{i,t}|$ and $|ar_{i,t}|$, which denote respectively the absolute value of adjusted ex-dividend stock returns and the absolute value of abnormal stock returns calculated as of equation (3.13).

4-2 Results

4-2-1 Lagged Cross Correlations

Tables 4-1 and 4-2 reports the correlation coefficients between abnormal turnover $AT^{k=20}$ and various SVI measures. For each pair of series, time lagged cross correlations are computed for δ = *-5*, *-4*, *… , 4*, *5*.

turnover variable: $AT^{k=20}$

	lag order(δ)										
	-5	-4	-3	-2	-1	0	1	2	3	4	5
Log($topic\ SVI^{**}$)	0,50%	0,86%	0,92%	1,21%	1,93%	3,13%	2,19%	1,32%	0,93%	0,62%	0,39%
$topic\ ASVI^{k=5}$	-4,33%	-3,82%	-3,99%	-2,17%	2,49%	10,04%	7,16%	4,18%	3,15%	2,32%	1,66%
$topic\ ASVI^{k=10}$	-3,14%	-1,07%	-0,29%	1,63%	5,81%	12,52%	8,96%	5,54%	4,26%	3,23%	2,48%
$topic\ ASVI^{k=20}$	0,35%	2,28%	2,84%	4,42%	8,01%	**13,91%**	10,00%	6,32%	4,77%	3,47%	2,50%
$topic\ ASVI^{k=40}$	1,60%	3,19%	3,52%	4,79%	7,89%	13,08%	9,27%	5,73%	4,17%	2,90%	1,93%
$topic\ ASVI^{k=60}$	1,75%	3,21%	3,48%	4,63%	7,51%	12,35%	8,71%	5,35%	3,86%	2,64%	1,72%

turnover variable: $AT^{k=20}$

	lag order(δ)										
	-5	-4	-3	-2	-1	0	1	2	3	4	5
Log($ticker\ SVI^{**}$)	0,29%	0,43%	0,46%	0,62%	0,96%	1,47%	0,92%	0,47%	0,28%	0,11%	0,11%
$ticker\ ASVI^{k=5}$	-2,60%	-2,32%	-2,19%	-0,52%	2,99%	7,98%	5,32%	3,16%	2,59%	2,04%	1,73%
$ticker\ ASVI^{k=10}$	-1,29%	0,05%	0,78%	2,40%	5,45%	9,60%	6,72%	4,40%	3,67%	2,97%	2,55%
$ticker\ ASVI^{k=20}$	1,57%	2,78%	3,30%	4,61%	7,07%	**10,41%**	7,42%	4,97%	3,99%	3,08%	2,52%
$ticker\ ASVI^{k=40}$	2,81%	3,70%	3,93%	4,85%	6,83%	9,62%	6,74%	4,40%	3,41%	2,50%	1,88%
$ticker\ ASVI^{k=60}$	2,60%	3,38%	3,54%	4,35%	6,15%	8,70%	5,93%	3,69%	2,73%	1,85%	1,24%

Tables 4-1 (above) and 4-2 (below): time lagged cross correlation coefficients between abnormal turnover and SVI variables. Choosing an observation window w=20 for both measures produces the best results. Correlations decrease as absolute lag increase, confirming the intuition that investors' attention leads to an increase in trading activity

As it was reasonable to expect, lags closer to 0 lead to higher correlation coefficients, while as the lag increases cross correlation fades, reaching negligible values for the extreme lag orders, occasionally dropping below 0. Series log(SVI) exhibits the poorest results among the series shown. Appendixes A2 and A3 report correlation values between and SVI and turnover series with different time windows *k*, all specifications produce results similar to tables 4-1 and 4-2.

Correlations between *AT* and *ASVI* computed using the same estimation window length seems to produce the best results. This can be intuitively explained by the fact that equal estimation windows capture comparable quantity of information.

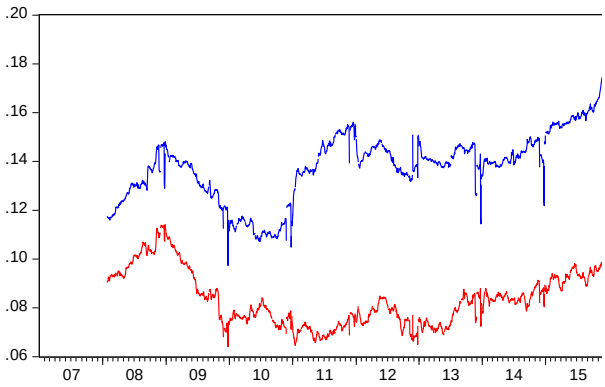
More interestingly, topic SVIs seem to be more strongly correlated to market activity than ticker SVIs. This seems to hold for all windows *k* and all lag orders δ displayed. However, it should be noted that results from table 4-1 and 4-2 are not fully comparable, because the number of cross sectional units available for the two SVI variables are very different. Therefore, the best performing indicators for both ticker types are selected ($ticker\ ASVI^{20}$ and $topic\ ASVI^{20}$), and the lagged correlations are computed on a reduced comparable sample that only includes those cross firms on which both indicators exhibit valid observations. As table 4-3 suggests, even in this reduced sample $topic\ ASVI$ dominates $ticker\ ASVI$. Figure 4-1 shows the moving correlation coefficient between $ASVI^{20}$ series and AT^{20}. As the figure indicates, topic SVI dominates ticker SVI for the whole sample period. Moreover, both moving correlations exhibit a peak during years 2007-2008, which could be explained by an increased attention towards the financial world due to the economic crisis.

Table 4-3: time-lagged cross correlation coefficients between abnormal turnover and SVI variables computed on a comparable sample. Sample used includes only those firms where both SVIs are available. In accordance with previous findings, *topic SVI* proves to be more correlated with market activity than *ticker SVI*

turnover variable: $AT^{k=20}$

lag order(δ)	$topic\ ASVI^{k=20}$	$ticker\ ASVI^{k=20}$
-5	0,5%	1,9%
-4	2,4%	3,1%
-3	3,2%	3,6%
-2	5,0%	4,9%
-1	8,8%	7,6%
0	14,6%	11,3%
1	10,2%	7,9%
2	6,7%	5,1%
3	5,3%	4,0%
4	3,7%	3,1%
5	2,7%	2,5%

Figure 4-1: Moving correlation cross-coefficient between abnormal turnover and ASVI variables. Correlation with topic SVI (blue line) dominates correlation with ticker SVI (red line) for the whole sample period.
Both correlation indicators seem to increase during periods of financial distress



Lastly, it is important to notice that, *ceteris paribus*, positive lag orders tend to produce higher cross-correlation coefficients. This holds regardless of the time window k chosen and for all the $\{Q, V\}$ pairs analyzed. A graphical representation of this phenomena is provided by figure 4-2, where the top performing indicators for ticker ASVI and topic ASVI are isolated and their measurements for positive and negative lag orders are overlaid. Positive lag orders, (connected with solid lines) always produce higher correlations than the correspondent negative lag orders (connected with dashed lines). This is particularly evident for topic SVI, and provides support to the hypothesis that SVI anticipates abnormal stock volume more than the opposite. This hypothesis is tested further in the next paragraph by conducting Granger-causality tests.

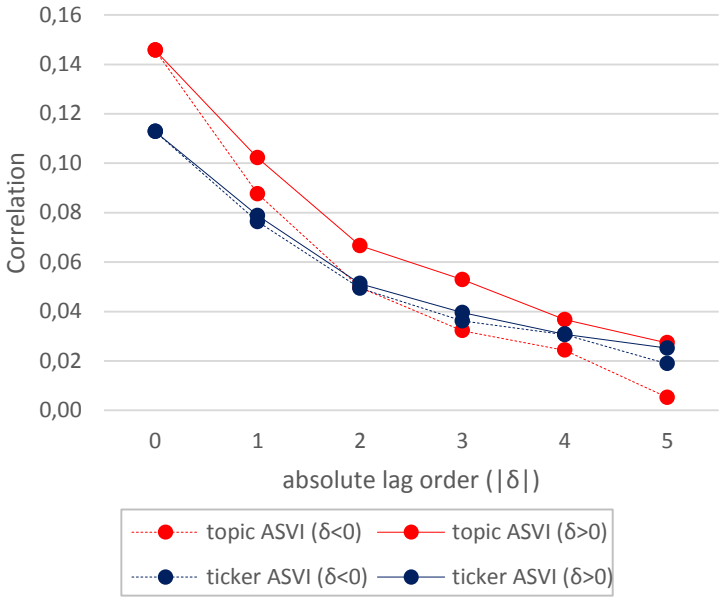


Figure 4-2: visual representation of correlation coefficients between abnormal turnover and ASVI. Positive and negative lag orders have been overlaid in order to highlight that correlations with positive lag orders (solid lines) dominate those with negative lag orders (dotted lines. This may suggest that current of values of SVI may anticipate future increases in stock turnover more than the opposite. ASVI measurement shown are computed with a window k=20

4-2-2 Granger-Causality test

Granger-causality test is run on each firm i, using time-series pairs $\{Q_i, V_i\}$. For each firm it is then derived the F-statistics of joint hypothesis (4.4 and 4.7) computed over fitted equations (4.3) and (4.6). It is also important to select an appropriate lag order: the most commonly used lag order selection criteria are based Akaike Information Criterion (AIC) and Hannan-Quinn Criterion (HQC). Optimal lag order is determined for each cross sectional unit. Because for almost all firms the optimal lag order is smaller than 4, Granger-causality test is performed for lag orders $l=1,2,3,4$.

Tables (4-4) and (4-5) summarize the results for selected $\{Q, V\}$ variable pairs, showing the number and percentage of firms for which each null hypothesis is rejected at different significance levels and for different lag orders. In almost all cases, Null hypothesis
$abnormal\ SVI\ does\ not\ Granger\ cause\ abnotmal\ turnover$ is rejected more often than Null hypothesis $abnormal\ turnover\ does\ not\ Granger\ cause\ abnormal\ SVI$. This test corroborates with previous results, providing further evidence that abnormal SVI values carry information that can predict future increases in trading activity. In addition, topic SVI seems to carry more predictive power than ticker SVI. This is consistent with the findings presented before by tables 4-1 and 4-2: *topic ASVI* appears to be more correlated with market activity than *ticker ASVI*.

Table (4-6) reports the outcome of pairwise Granger causality tests and stock volatility proxies. Absolute returns measure the full volatility of the stock, while absolute abnormal returns exclude market volatility, providing a measure of idiosyncratic risk. In accordance with previous results of this study, *topic SVI* seems to be a better predictor of stock volatility than *ticker SVI,* suggesting that the latter is a rather noisy indicator. Moreover, in most cases SVI seems to Granger-cause absolute abnormal returns more often than absolute returns. This suggests that SVI is not simply an indicator of investment sentiment towards the market as a whole, but it also conveys information regarding the specific firm's risk.

	Lag order:1			Lag order:2			Lag order:3			Lag order:4		
Causality	**p = 0,1**	**p = 0,05**	**p = 0,01**	**p = 0,1**	**p = 0,05**	**p = 0,01**	**p = 0,1**	**p = 0,05**	**p = 0,01**	**p = 0,1**	**p = 0,05**	**p = 0,01**
$ticker\ ASVI^{k=20} \rightarrow AT^{k=20}$	54 (44,3%)	41 (33,6%)	30 (24,6%)	51 (41,8%)	43 (35,2%)	29 (23,8%)	53 (43,4%)	46 (37,7%)	28 (23,0%)	54 (44,3%)	41 (33,6%)	27 (22,1%)
$AT^{k=20} \rightarrow ticker\ ASVI^{k=20}$	45 (36,9%)	34 (27,9%)	23 (18,9%)	50 (41,0%)	41 (33,6%)	18 (14,8%)	54 (44,3%)	35 (28,7%)	21 (17,2%)	53 (43,4%)	35 (28,7%)	20 (16,4%)
Both	24 (19,7%)	16 (13,1%)	8 (6,6%)	25 (20,5%)	19 (15,6%)	5 (4,1%)	25 (20,5%)	17 (13,9%)	8 (6,6%)	25 (20,5%)	19 (15,6%)	6 (4,9%)
$ticker\ ASVI^{k=40} \rightarrow AT^{k=40}$	50 (41,0%)	39 (32,0%)	24 (19,7%)	47 (38,5%)	40 (32,8%)	26 (21,3%)	51 (41,8%)	43 (35,2%)	25 (20,5%)	54 (44,3%)	40 (32,8%)	24 (19,7%)
$AT^{k=40} \rightarrow ticker\ ASVI^{k=40}$	40 (32,8%)	31 (25,4%)	18 (14,8%)	53 (43,4%)	35 (28,7%)	19 (15,6%)	56 (45,9%)	39 (32,0%)	18 (14,8%)	50 (41,0%)	34 (27,9%)	20 (16,4%)
Both	21 (17,2%)	13 (10,7%)	6 (4,9%)	22 (18,0%)	12 (9,8%)	5 (4,1%)	21 (17,2%)	14 (11,5%)	5 (4,1%)	24 (19,7%)	11 (9,0%)	8 (6,6%)
	Lag order:1			**Lag order:2**			**Lag order:3**			**Lag order:4**		
Causality	**p = 0,1**	**p = 0,05**	**p = 0,01**	**p = 0,1**	**p = 0,05**	**p = 0,01**	**p = 0,1**	**p = 0,05**	**p = 0,01**	**p = 0,1**	**p = 0,05**	**p = 0,01**
$topic\ ASVI^{k=20} \rightarrow AT^{k=20}$	249 (60,6%)	220 (53,5%)	145 (35,3%)	241 (58,6%)	194 (47,2%)	117 (28,5%)	234 (56,9%)	187 (45,5%)	123 (29,9%)	225 (54,7%)	176 (42,8%)	110 (26,8%)
$AT^{k=20} \rightarrow topic\ ASVI^{k=20}$	230 (56,0%)	204 (49,6%)	141 (34,3%)	222 (54,0%)	176 (42,8%)	113 (27,5%)	220 (53,5%)	180 (43,8%)	111 (27,0%)	224 (54,5%)	181 (44,0%)	121 (29,4%)
Both	159 (38,7%)	130 (31,6%)	63 (15,3%)	140 (34,1%)	88 (21,4%)	36 (8,8%)	137 (33,3%)	89 (21,7%)	37 (9,0%)	130 (31,6%)	85 (20,7%)	39 (9,5%)
$topic\ ASVI^{k=40} \rightarrow AT^{k=40}$	242 (58,9%)	205 (49,9%)	153 (37,2%)	244 (59,4%)	202 (49,1%)	132 (32,1%)	237 (57,7%)	198 (48,2%)	127 (30,9%)	228 (55,5%)	181 (44,0%)	119 (29,0%)
$AT^{k=40} \rightarrow topic\ ASVI^{k=40}$	211 (51,3%)	174 (42,3%)	105 (25,5%)	191 (46,5%)	146 (35,5%)	82 (20,0%)	205 (49,9%)	160 (38,9%)	92 (22,4%)	210 (51,1%)	175 (42,6%)	111 (27,0%)
Both	142 (34,5%)	110 (26,8%)	61 (14,8%)	121 (29,4%)	80 (19,5%)	27 (6,6%)	134 (32,6%)	87 (21,2%)	32 (7,8%)	124 (30,2%)	86 (20,9%)	41 (10,0%)

Tables 4-4 (above) and 4-5 (below): results of pairwise Granger-causality test. For each firm a pair of series is used, with one series measuring SVI and the other series measuring trading activity. The sample covers period 1/1/2007-31/11/2015. The number of cross sectional units depends on the SVI availability. The figures represent the number of firms for which the null hypothesis is rejected at level of confidence p=10%,5%,1%. The numbers in brackets represent the percentage of firms for which the null hypothesis is rejected. As an example, the top leftmost figure indicates that at lag order l=1 and at significance level p of 10%, the null hypothesis that ticker SVI does not Granger-cause abnormal turnover is rejected 54 times, which correspond to 44,3% of the total number of firms in sample (122). Label "both" refers to the number of firms for witch the null hypothesis is rejected on both directions, i.e. the firms where S→V and V→S.

Causality	Lag order:1			Lag order:2			Lag order:3			Lag order:4		
	p = 0,1	p = 0,05	p = 0,01	p = 0,1	p = 0,05	p = 0,01	p = 0,1	p = 0,05	p = 0,01	p = 0,1	p = 0,05	p = 0,01
$ticker\ ASVI^{k=20} \rightarrow \lvert ar_{i,t} \rvert$	28 (23,0%)	20 (16,4%)	9 (7,4%)	25 (20,5%)	16 (13,1%)	8 (6,6%)	23 (18,9%)	17 (13,9%)	10 (8,2%)	26 (21,3%)	19 (15,6%)	10 (8,2%)
$ticker\ ASVI^{k=20} \rightarrow \lvert r_{i,t} \rvert$	27 (22,1%)	21 (17,2%)	8 (6,6%)	33 (27,0%)	18 (14,8%)	7 (5,7%)	28 (23,0%)	15 (12,3%)	6 (4,9%)	25 (20,5%)	18 (14,8%)	7 (5,7%)
$ticker\ ASVI^{k=40} \rightarrow \lvert ar_{i,t} \rvert$	35 (28,7%)	27 (22,1%)	9 (7,4%)	26 (21,3%)	19 (15,6%)	8 (6,6%)	24 (19,7%)	20 (16,4%)	10 (8,2%)	27 (22,1%)	18 (14,8%)	10 (8,2%)
$ticker\ ASVI^{k=40} \rightarrow \lvert r_{i,t} \rvert$	38 (31,1%)	29 (23,8%)	15 (12,3%)	32 (26,2%)	27 (22,1%)	10 (8,2%)	30 (24,6%)	22 (18,0%)	8 (6,6%)	27 (22,1%)	20 (16,4%)	8 (6,6%)
$topic\ ASVI^{k=20} \rightarrow \lvert ar_{i,t} \rvert$	124 (30,4%)	94 (23,0%)	54 (13,2%)	106 (26,0%)	79 (19,4%)	36 (8,8%)	105 (25,7%)	73 (17,9%)	48 (11,8%)	108 (26,5%)	83 (20,3%)	46 (11,3%)
$topic\ ASVI^{k=20} \rightarrow \lvert r_{i,t} \rvert$	87 (21,2%)	65 (15,8%)	33 (8,0%)	86 (20,9%)	53 (12,9%)	19 (4,6%)	92 (22,4%)	57 (13,9%)	22 (5,4%)	100 (24,3%)	69 (16,8%)	34 (8,3%)
$topic\ ASVI^{k=40} \rightarrow \lvert ar_{i,t} \rvert$	136 (33,3%)	108 (26,5%)	73 (17,9%)	124 (30,4%)	96 (23,5%)	47 (11,5%)	117 (28,7%)	86 (21,1%)	45 (11,0%)	116 (28,4%)	92 (22,5%)	49 (12,0%)
$topic\ ASVI^{k=40} \rightarrow \lvert r_{i,t} \rvert$	114 (27,7%)	87 (21,2%)	46 (11,2%)	107 (26,0%)	73 (17,8%)	35 (8,5%)	102 (24,8%)	68 (16,5%)	34 (8,3%)	105 (25,5%)	66 (16,1%)	36 (8,8%)

Table 4-6 results of pairwise Granger-causality test. For each firm a pair of series is used, with one series measuring SVI and the other series measuring stock volatility. The sample covers period 1/1/2007-31/11/2015. The number of cross sectional units (firms) depends on the SVI availability (411 for ticker SVI and 122 for topic SVI) Lag orders l=1,2,3,4 are used. The figures represent the number of firms for which the null hypothesis is rejected at level of confidence p=10%,5%,1%. The numbers in brackets represent the percentage of firms for which the null hypothesis is rejected. Topic SVI is generally a better predictor of future stock volatility than ticker SVI. More interestingly, SVI predicts abnormal returns better than ex-dividend returns

Chapter 5

SVI and stock returns

This section examines whether ASVI carries information that anticipate future abnormal stock returns. Chapter 4 has highlighted that there is a correlation between SVI and market activity. Therefore, it is not unreasonable to expect this to translate in some sort of stock price behavior. As pointed out by (Bank, Larch, and Peter 2011), academic literature offers two contrasting hypotheses regarding investor attention: (Merton 1987) suggests that stocks subject to lower levels of investor attention can provide higher returns in order to compensate for idiosyncratic risk, which cannot be diversified. However, (Barber et al. 2008) posit that stock with higher search volumes should exhibit higher returns in virtue of the fact that attention affects more the buying side than the selling side. In fact, they argue that investors willing to buy have the option to choose among a wide variety of stocks and will therefore generate a higher number of queries than the investors willing to sell, who only have a limited choice. Most of the scientific literature up to date seems to confirm this theory; i.e. increased SVI values are followed by higher stock returns. Da et al. (2011) use Google Trends weekly data for US Russel 3000 firms in order to show that a positive abnormal SVI is followed by higher returns in the subsequent 2 weeks; Bank, Larch, and Peter (2011) conduct a similar analysis on the German market, which results in similar results. Joseph, Babajide Wintoki, and Zhang (2011) find evidence that further confirms SVI's influence over abnormal returns, showing also that the intensity of this correlation is affected by the liquidity of the stock traded. However, there seems to be a lack of studies that test whether these effects are also present when daily observations are used. Intuition would suggest that using daily observations could potentially lead to improved results, because daily SVI is a more timely indicator than weekly SVI and allows investors to react more quickly to new information being released. In order to test whether this hypothesis is true, the following approach is pursued: first, a series of cross sectional linear regressions is run in order to simulate a security selection trading strategy and test the correlation between ASVI and returns. The average regression coefficient, measured in accordance with Fama-Macbeth method, provides insights regarding the profitability of the trading strategy being tested. Therefore, the regression coefficient is computed multiple times, using different dependent variables (returns) and regressors (ASVI specifications). For each regression, the statistical significance of the coefficient is measured via a t-test in order to assess the reliability of the results.

Then, using an approach similar to Joseph, Babajide Wintoki, and Zhang (2011), alternative trading strategies based on sorted portfolios are simulated and their abnormal returns are computed by controlling for the risk factors of the most commonly used index models. The rest of this chapter describes the methodology, followed by results obtained.

5-1 Metodology

5-1-2 Rolling regressions

Pedersen (2015) provides a theoretical justification of the approach that will be described in this section by showing that "*a cross sectional regression corresponds to a security selection strategy*". In line with his approach, for each time period *t*, the following regression is computed:

$$R_t^i = a + bQ_{t-l}^i + \varepsilon_i^t \tag{5.1}$$

Where R_t^i denotes the return of stock i and Q_{t-l}^i represents a lagged search query variable (lag order *l)*. Therefore, the estimated regression coefficient $\hat{b}_t$ for each period *t* will be:

$$\hat{b}_t = \frac{\sum_i (Q_t^i - \overline{Q_t}) R_{t+l}^i}{\sum_i (Q_t^i - \overline{Q_t})^2} = \sum_i x_t^i R_{t+l}^i \tag{5.2}$$

Pedersen (2015) shows that $\hat{b}_t$ can also be interpreted as the profit of a long-short strategy, where the position for each security is:

$$x_t^i = \frac{(Q_t^i - \overline{Q_t})}{\sum_i (Q_t^i - \overline{Q_t})^2} \tag{5.3}$$

In other words, if $\hat{b}_t$ is positive, a portfolio that is long on stocks with high SVI and short on stocks with low SVI would have achieved positive returns. In practice, this weighting is not the most commonly used approach in trading simulation; nevertheless, the sign of the coefficient is still an important indicator,

because it suggests the sign of the relationship between ASVI and returns. More importantly, $\hat{b}_t$ can be used to compute the overall estimate of the coefficient, that using the Fama-Macbeth (Fama and MacBeth, 1973) method is given by:

$$\hat{b} = \frac{1}{T}\sum_{t=1}^{T}\hat{b}_t \tag{5.4}$$

As the equation suggests, $\hat{b}$ represents the relationship between Q and returns over the entire sample period. If $\hat{b}$ is positive and significant, it can be said that an increase of Q_{t-l} anticipates an increase in returns R_t

It is also necessary to test the significance of said coefficient. Therefore, the volatility of the coefficient is computed:

$$\hat{\sigma} = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(\hat{b}_t - \hat{b})^2} \tag{5.5}$$

And the t-statistic is derived as follows:

$$t - statistic = \sqrt{T}\frac{\hat{b}}{\hat{\sigma}} \tag{5.6}$$

In order for the trading strategy to be profitable and reliable, a t-statistic with high absolute value is desirable. Regression (5.1) will be computed several times using different dependent variables (returns) and regressors (SVI measures).

5-1-3 Sorted portfolios

The procedure described in the previous paragraph can be used to produce a preliminary assessment pertaining potential applications of ASVI as a trading signal. This section introduces an alternative and more realistic approach to the problem, based on sorted portfolios. This method will be used to back-test different trading strategy specifications.

Within this study's framework, a sorted-portfolio trading strategy of lag order *l* can be described as follows: every day *t*, the firms composing the sample are sorted into five quintiles based on the ASVI registered on the day *t-l*. These quintiles determine the equally weighted components of five stock portfolios: P1 contains the firms with low ASVI, while P5 is the high ASVI portfolio. One additional portfolio is computed: P5-P1, which is a self-financing portfolio that goes short on low ASVI stocks and long on high ASVI stocks. Portfolios are rebalanced daily in order to track ASVI changes. Portfolios' excess returns are then regressed on risk factors in order to determine abnormal returns. In order to ensure the robustness of the results, risk factors from three different market models will be used: CAPM , Fama-French 3 factor model (Fama and French 1993) and (Carhart 2016) four factor model. More formally, the abnormal returns of each market model are obtained as follows:

$$R_P - R_f = \alpha + \beta_1(R_M - R_F) + \varepsilon_t \tag{5.7}$$

$$R_P - R_f = \alpha + \beta_1(R_M - R_F) + \beta_2 SMB_t + \beta_3 HML_t + \varepsilon_t \tag{5.8}$$

$$R_P - R_f = \alpha + \beta_1(R_M - R_F) + \beta_2 SMB_t + \beta_3 HML_t + \beta_4 UMD_t + \varepsilon_t \tag{5.91}$$

Where α denotes the return component of the trading strategy that is not explained by the model, i.e. the active return of the portfolio. In line with existing literature, it is reasonable to expect α to be higher for portfolio with high SVI. Therefore, a positive and significant α on the P5-P1 portfolio would indicate the profitability of a strategy that goes long on high SVI stocks and short on low SVI stocks.

In addition to the regressions above, portfolios are evaluated on the basis of the Sharpe Ratio, computed as:

$$SR = \frac{R_P - R_f}{\sigma_P} \tag{5.10}$$

That is the ratio of portfolio's excess returns over portfolio's volatility. This measure provides a risk-adjusted measure of the returns.

5-1-4 Simulated trading strategies

The sorted-portfolio approach is adopted in order to test three different type of trading strategies that differ in terms of lag order selected and assumptions made:

Trading strategy I - "Cheating"

Under this approach, it is assumed that the investor trades on day *t* based on ASVI changes realized on the same day *t.* Clearly, this approach is very biased and unrealistic: ASVI observations for time *t* refer to all searches computed until the end of the day (midnight). Therefore, when the investor trades on day *t*, the $ASVI_t$ specification is not yet available. Nevertheless, assessing this strategy is relevant from a theoretical perspective: it suggests if ASVI is correlated with positive returns and it provides insights on the magnitude of such relation. Moreover, comparing this approach to the following ones allows to observe how quickly the markets align with new information being released

Trading strategy II - Semi-Biased

A more realistic approach would require the investor to trade on *t* after observing ASVI changes on *t-1*, measuring the profitability of the strategy based on end-of-day returns achieved (r_t) . This approach is more realistic than the previous one, because the value of $ASVI_{t-1}$ is available on *t*. However, it still includes a bias component: the profitability of the trading strategy is assessed by measuring end-of-day returns computed as $r_t = (P_t^{Close}/P_{t-1}^{Close}) - 1$. As the calculation suggests, these returns are based on stock closing prices on time *t* and *t-1*. Therefore, this strategy implicitly assumes the investor can buy at time *t* stocks at closing price P_{t-1}^{Close}. This is clearly unrealistic, because stocks are traded after hours and their price can therefore change in relation to information that may be released when the market is closed. Therefore, this strategy is surely less biased than the previous one, but it is still prone to *look-ahead bias.*

Trading strategy III - Open-to-open

In correct the bias of previous strategy, r_t must be replaced by a more suitable indicator: using open-to open returns (r^{open}) is an appropriate solution. Open-to-open returns are computed using stock opening prices as a reference ($r_t^{Open} = (P_t^{Open}/P_{t-1}^{Open}) - 1$) . Therefore, measuring the strategy profitability via r^{open} is equivalent to assume that investor observes $ASVI_{t-1}$ at the end of day *t-1,* and use such information to trade when the market opens on day *t.* After holding the stock for 24 hours, the return registered will be $r_t^{Open} = P_{t+1}^{Open}/P_t^{Open} - 1)$. This strategy is not biased because when investor trades he only uses information that is already available. It should also be noted that computing abnormal returns of this strategy requires a modification of equations (5.7) to (5.9). This is due to the fact that risk factors are provided in a form that is supposed to be used with end-of-day returns. In order to correct for this distortion, regressions are adjusted by included the lagged terms of the risk factors for each model.

$$R_{P,t}^{open} - R_f = \alpha + \sum_{i=0}^{1} \beta_{1,i}\left(R_{M,t-i} - R_F\right) + \varepsilon_t \tag{5.11}$$

$$R_{P,t}^{open} - R_f = \alpha + \sum_{i=0}^{1} \beta_{1,i}\left(R_{M,t-i} - R_F\right) + \sum_{i=0}^{1} \beta_{2,i} SMB_{t-i} + \sum_{i=0}^{1} \beta_{3,i} HML_{t-i} + \varepsilon_t \tag{5.12}$$

$$R_{P,t}^{open} - R_f = \alpha + \sum_{i=0}^{1} \beta_{1,i}\left(R_{M,t-i} - R_F\right) + \sum_{i=0}^{1} \beta_{2,i} SMB_{t-i} + \sum_{i=0}^{1} \beta_{3,i} HML_{t-i} + \sum_{i=0}^{1} \beta_{3,i} UMD_{t-i} + \varepsilon_t \tag{5.13}$$

The overall contribution of each risk factor on the portfolio returns is computed by summing the betas referred to each type of risk factor. (e.g in equation (5.12), total exposure to SMB is given by $\beta_{2,0} + \beta_{2,1}$)

5-2 Results

5-2-1 Rolling regressions

Table 5-1 presents the coefficients obtained via equation 5.1 and their t-statistics. Two different ASVI variables have been used as regressors: one for *topic SVI* and one for *ticker SVI*. Independent variables used are: ex-dividend returns, abnormal returns computed as of equation 3.13, open-to-open ex-dividend returns. Several insights can be drawn from these results: in accordance to previous literature, all coefficients exhibit positive sign, suggesting that a SVI increase is correlated with an increase in returns. As it was reasonable to expect, the most statistically significant coefficients are registered on regressions where *l=0*, while for *l=1* the coefficients are less significant, which indicates that the market has already incorporated at least part of the information that $ASVI_{t-1}$ carried. For both lag specifications, *topic ASVI* seems to outperform *ticker ASVI*, suggesting that a trading strategy based on *ticker ASVI* is unlikely to be profitable. *Topic SVI* regression coefficients are more robust and produce relatively high t-stats when regressed over returns and abnormal returns. However, both ASVI indicators do not seem to perform well when paired with open-to-open returns. This is not surprising: as it was mentioned in the previous section, the period on which r_{t-1}^{Open} is realized does not overlap at all with the time interval on which $ASVI_t$ is computed. The same cannot be said for r and ar. For this reason, it can be inferred that coefficients obtained with r^{open} are lower because the trading strategy that the regression simulates is not biased, while the other regressions contain some *look-ahead bias component.*

Table 5-1: regression coefficients and t-stats of several specifications of equation 5.1

		Lag: 1		Lag: 0	
regressor	**indep. variable**	**coeff**	**(t-stat)**	**coeff**	**(t-stat)**
$topic\ ASVI^{k=40}$	r	0,00183	(1,41)	0,00668	(3,98)
$topic\ ASVI^{k=40}$	ar	0,00249	(1,58)	0,00810	(3,80)
$topic\ ASVI^{k=40}$	r^{open}	0,00018	(0,13)	0,00315	(2,02)
$ticker\ ASVI^{k=40}$	r	0,00155	(0,71)	0,00551	(2,16)
$ticker\ ASVI^{k=40}$	ar	0,00011	(0,04)	0,00579	(1,76)
$ticker\ ASVI^{k=40}$	r^{open}	0,00019	(0,10)	0,00348	(1,53)

5-2-2 Sorted-Portfolios

Table 5-2 presents the results of *Trading Strategy I - "Cheating"*. The SVI signal used to sort the portfolios is $topic\ ASVI^{k=40}$. The table reports factor loadings computed for all three market models examined. Coefficients α indicate the average daily abnormal return, that is the part of returns which is not explained by the model. In order to make this figure more readable, annualized abnormal returns are also computed: $ar^{annual} = (1 + \alpha)^{252} - 1$.

There is a nearly monotonic relation between abnormal returns and ASVI intensity: high SVI portfolios achieve significantly higher abnormal returns. This result is consistent across all three market models used, and it corroborates the result of previous section. Even though the pattern exhibited by returns is very clear, there is not a distinct pattern that emerges from other risk factors. P5 exhibits higher market exposure than all other portfolios, but there seem not to be a clear trend among P1-2-3-4 under this perspective. Nevertheless, the profitability of high SVI portfolios is clear, and it is also confirmed by the figures presented in table 5-3, where average excess returns and Sharpe ratios are computed. It can easily be seen that high SVI portfolios score better in terms of Sharpe Ratio, suggesting better risk-adjusted returns. Appendix A3 offers a graphical representation of cumulated excess returns registered by the portfolios.

The results presented so far provide evidence that there is a significant correlation between high SVI and positive abnormal returns. However, *Trading Strategy I* does not indicate whether this holds when trading activity is deferred, i.e. when the *look-ahead bias* is reduced. The analysis of *Trading Strategy II - "Semi-Biased"* can provide relevant insights in this sense. As table 5-4 illustrates, high SVI portfolios are still characterized by higher abnormal returns, but the effect is not as strong as shown in *Trading Strategy I*. The relatively high abnormal returns achieved by P3 contradicts the nearly monotonic relation between SVI and

returns that was highlighted before. Nevertheless, the intercept of P5-1 is still positive and significant at a 10% confidence level. As table 5-5 illustrates, high SVI portfolio P5 is more risky than the others, but the increase in volatility is more than compensated by superior returns, resulting in a higher Sharpe Ratio. Overall, the results appear to be in line with existing literature described in section 2-2. Appendix A4 offers a graphical representation of cumulated excess returns registered by the portfolios.

Rolling regressions discussed in section 5-2-1 showed that using open-to-open returns as independent variable did not produce statistically significant coefficients. Moreover, results aforementioned in this section suggest that deferring trading activity hinders the profitability of the trading strategy. Based on these premises, it is reasonable to expect *Trading Strategy III – Open-to-Close* to perform worse than the others; this intuition is confirmed by table 5-6. Low SVI portfolio is still the worst performing of the group, but there seem to be no clear pattern that can describe the returns of other portfolios, and none of the portfolio exhibits significantly positive abnormal returns. No additional insights can be drawn from table 5-7, where there seem to be no clear correlation between SVI intensity and realized Sharpe Ratio. As it was explained, this strategy assumes that the investor trades when market opens on the basis of ASVI information of the previous day ;i.e. the investor acts only a few hours after the SVI is available. The results presented seem to suggest that these few hours are enough to significantly reduce the profitability of the trading strategy.

As a last note, it should be pointed out that transaction costs have not been considered in the analysis. This constitutes a further limitation to the findings, especially in consideration of the fact that, because the portfolios are rebalanced daily, transaction costs should expected to be significant.

The analysis just discussed has also be conducted using *ticker SVI* as a signal for portfolio sorting. In consistency with results reported by table 5-1, *ticker SVI* proved to be a poorer indicator of investor attention and did not return significant results. For the sake of brevity, and due to the lack of importance of the results produced, output for *ticker SVI* is not reported.

Table 5-2: Analysis of trading strategy I – "Cheating". *,** and *** denote significance at 10%, 5% and 1% respectively

	ar^{annual}	α	MKT	SMB	HML	UMD	R^2
CAPM							
P1	-1,994%	-0,0001	1,1133***				0,960
		(-1,21)	(231,29)				
P2	0,321%	0,0000	1,0993***				0,965
		(0,21)	(246,58)				
P3	1,825%	0,0001	1,0920***				0,965
		(1,18)	(247,01)				
P4	5,099%	0,0002 ***	1,1056***				0,964
		(3,16)	(243,29)				
P5	6,248%	0,0002 ***	1,1532***				0,950
		(3,09)	(203,65)				
P5-1	8,409%	0,0003 ***	0,0398***				0,019
		(3,85)	(6,58)				
3FF							
P1	-1,612%	-0,0001	1,0836 ***	0,0897 ***	0,1173 ***		0,963
		(-1,01)	(207,60)	(8,35)	(10,62)		
P2	0,877%	0,0000	1,0652 ***	0,0587 ***	0,1530 ***		0,969
		(0,60)	(224,31)	(6,01)	(15,22)		
P3	2,406%	0,0001 *	1,0562 ***	0,0659 ***	0,1586 ***		0,969
		(1,65)	(225,75)	(6,85)	(16,02)		
P4	5,561%	0,0002 ***	1,0748 ***	0,0778 ***	0,1276 ***		0,967
		(3,58)	(219,39)	(7,72)	(12,31)		
P5	6,881%	0,0003 ***	1,1173 ***	0,0572 ***	0,1632 ***		0,953
		(3,52)	(182,07)	(4,53)	(12,56)		
P5-1	8,631%	0,0003 ***	0,0337 ***	-0,0324 **	0,0459 ***		0,027
		(3,96)	(4,96)	(-2,32)	(3,19)		
3FF+UMD							
P1	-1,478%	-0,0001	1,0705 ***	0,0905 ***	0,0422 ***	-0,0848 ***	0,965
		(-0,95)	(206,18)	(8,68)	(3,37)	(-11,60)	
P2	1,048%	0,0000	1,0489 ***	0,0597 ***	0,0594 ***	-0,1057 ***	0,972
		(0,75)	(228,25)	(6,47)	(5,36)	(-16,33)	
P3	2,597%	0,0001 *	1,0382 ***	0,0670 ***	0,0555 ***	-0,1165 ***	0,973
		(1,91)	(232,90)	(7,48)	(5,16)	(-18,55)	
P4	5,737%	0,0002 ***	1,0587 ***	0,0788 ***	0,0350 ***	-0,1046 ***	0,970
		(3,89)	(222,24)	(8,23)	(3,04)	(-15,59)	
P5	7,038%	0,0003 ***	1,1031 ***	0,0581 ***	0,0818 ***	-0,0919 ***	0,955
		(3,69)	(179,88)	(4,72)	(5,53)	(-10,63)	
P5-1	8,643%	0,0003 ***	0,0326 ***	-0,0324 **	0,0397 **	-0,0070	0,027
		(3,96)	(4,69)	(-2,32)	(2,36)	(-0,72)	

Table 5-3: Realized profits and Sharp Ratios for trating Strategy I – "Cheating"

	P1	P2	P3	P4	P5	P5-1
Av. Ret	0,000312	0,000400	0,000456	0,000587	0,000647	0,000335
Volatility	0,015602	0,015369	0,015265	0,015464	0,016253	0,003943
Sharpe R.	0,020013	0,026024	0,029900	0,037947	0,039793	0,084833

Table 5-4: Analysis of trading strategy II – "Semi-Biased". *,** and *** denote significance at 10%, 5% and 1% respectively

	ar^{annual}	α	MKT	SMB	HML	UMD	R^2
CAPM							
P1	0,329%	0,0000	1,1281***				0,958
		(0,19)	(224,52)				
P2	3,310%	0,0001 **	1,0878***				0,961
		(2,02)	(233,81)				
P3	0,871%	0,0000	1,0915***				0,965
		(0,56)	(244,65)				
P4	3,732%	0,0001 **	1,1108***				0,962
		(2,25)	(236,08)				
P5	3,810%	0,0001 **	1,1472***				0,954
		(2,01)	(213,30)				
P5-1	3,470%	0,0001 *	0,0192***				0,005
		(1,66)	(3,22)				
3FF							
P1	0,645%	0,0000	1,1036 ***	0,0762 ***	0,0957 ***		0,960
		(0,38)	(199,80)	(6,70)	(8,18)		
P2	3,889%	0,0002 **	1,0521 ***	0,0693 ***	0,1569 ***		0,965
		(2,50)	(212,29)	(6,79)	(14,96)		
P3	1,376%	0,0001	1,0576 ***	0,0789 ***	0,1432 ***		0,968
		(0,93)	(222,29)	(8,06)	(14,22)		
P4	4,372%	0,0002 ***	1,0718 ***	0,0735 ***	0,1718 ***		0,967
		(2,80)	(215,87)	(7,19)	(16,35)		
P5	4,394%	0,0002 **	1,1138 ***	0,0491 ***	0,1535 ***		0,957
		(2,39)	(190,85)	(4,09)	(12,42)		
P5-1	3,725%	0,0001 *	0,0102	-0,0270 **	0,0578 ***		0,015
		(1,78)	(1,54)	(-1,97)	(4,10)		
3FF+UMD							
P1	0,786%	0,0000	1,0901 ***	0,0770 ***	0,0185	-0,0871 ***	0,962
		(0,47)	(198,07)	(6,96)	(1,39)	(-11,24)	
P2	4,056%	0,0002 ***	1,0366 ***	0,0702 ***	0,0680 ***	-0,1004 ***	0,968
		(2,73)	(213,93)	(7,21)	(5,82)	(-14,71)	
P3	1,556%	0,0001	1,0404 ***	0,0800 ***	0,0444 ***	-0,1115 ***	0,972
		(1,12)	(227,49)	(8,70)	(4,03)	(-17,31)	
P4	4,551%	0,0002 ***	1,0553 ***	0,0745 ***	0,0769 ***	-0,1072 ***	0,970
		(3,07)	(218,86)	(7,69)	(6,61)	(-15,79)	
P5	4,560%	0,0002 **	1,0985 ***	0,0501 ***	0,0657 ***	-0,0991 ***	0,960
		(2,56)	(189,77)	(4,31)	(4,70)	(-12,15)	
P5-1	3,745%	0,0001 *	0,0084	-0,0269 **	0,0472 ***	-0,0119	0,016
		(1,79)	(1,23)	(-1,97)	(2,87)	(-1,24)	

Table 5-5: Realized profits and Sharp Ratios for trating Strategy II – "Semi-Biased"

	P1	P2	P3	P4	P5	P5-1
Av. Ret	0,00041	0,00051	0,00042	0,00054	0,00055	0,00014
Volatility	0,01583	0,01524	0,01526	0,01555	0,01613	0,00384
Sharpe R.	0,02593	0,03363	0,02744	0,03451	0,03425	0,03700

Table 5-6: Analysis of trading strategy III – "Open-to-Open". *,** and *** denote significance at 10%, 5% and 1% respectively

	ar^{annual}	α	MKT	SMB	HML	UMD	R^2
CAPM							
P1	0,448%	0,0000	1,1321				0,557
		(0,08)					
P2	3,804%	0,0001	1,1122				0,544
		(0,69)					
P3	1,236%	0,0000	1,0961				0,536
		(0,22)					
P4	3,381%	0,0001	1,1106				0,547
		(0,61)					
P5	1,841%	0,0001	1,1733				0,544
		(0,32)					
P5-1	1,387%	0,0001	0,0412				0,015
		(0,64)					
3FF							
P1	1,070%	0,0000	1,0871	0,0906	0,1433		0,575
		(0,20)					
P2	4,463%	0,0002	1,0715	0,0573	0,1441		0,564
		(0,82)					
P3	1,909%	0,0001	1,0534	0,0581	0,1488		0,559
		(0,35)					
P4	4,005%	0,0002	1,0725	0,0480	0,1328		0,570
		(0,74)					
P5	2,586%	0,0001	1,1285	0,0482	0,1617		0,570
		(0,46)					
P5-1	1,500%	0,0001	0,0414	-0,0424	0,0183		0,045
		(0,70)					
3FF+UMD							
P1	1,263%	0,0000	1,0678	0,0962	0,0312	-0,1258	0,580
		(0,24)					
P2	4,670%	0,0002	1,0515	0,0632	0,0280	-0,1302	0,570
		(0,86)					
P3	2,115%	0,0001	1,0331	0,0644	0,0292	-0,1350	0,566
		(0,40)					
P4	4,203%	0,0002	1,0532	0,0538	0,0207	-0,1261	0,575
		(0,78)					
P5	2,756%	0,0001	1,1120	0,0535	0,0632	-0,1117	0,574
		(0,49)					
P5-1	1,474%	0,0001	0,0442	-0,0427	0,0319	0,0141	0,046

Table 5-7: Realized profits and Sharp Ratios for trating Strategy III – "Open-to-Open"

	P1	P2	P3	P4	P5	P5-1
Av. Ret	0,00041	0,00054	0,00043	0,00052	0,00048	0,00007
Volatility	0,01518	0,01495	0,01491	0,01501	0,01570	0,00405
Sharpe R.	0,02729	0,03597	0,02902	0,03470	0,03080	0,01707

Chapter 6

SVI and implied volatility

Chapter 4 highlighted that search query volumes carry information regarding future levels of trading activity: an increase in investors information demand is followed by increased trading activity. Trading activity is notoriously correlated with stock volatility. Therefore, it reasonable to expect search query to be correlated stock volatility. The existence of such correlation has been shown by (Dimpfl and Jank 2012) and (cit), who demonstrate that several volatility forecasting models can be improved by including search queries among the regressors.

However, very little research has been conducted in order to investigate whether search queries can anticipate the expectations regarding future stock volatility, whose most common measure is *implied volatility*, and its theoretical justification lies upon Black–Scholes–Merton option pricing model (cit.). Black-Scholes show that the current price of a call option C_0 for a non-dividend paying stock can be written as:

$$C_0 = S_0 N(d_1) - X e^{-rT} N(d_2) \tag{6.1}$$

Where

$$d_1 = \frac{\ln\left({S_0}/{X}\right) + \left(r + {\sigma^2}/{2}\right)T}{\sigma\sqrt{T}}$$

$$d_2 = d_1 - \sigma\sqrt{T}$$

Where S_0 ist he current stock price, $N(d)$ is the cumulative distribution function of the standard normal distribution, X is the exercise price, r indicates the risk-free interest rate (annualized and continuously compounded), T is the time to expiration and σ represents the standard deviation of the annualized continuously compounded rate of return.

Given the formula above, it is possible to derive the price of a put option via put-call parity principle (cit.); i.e. the price of a Put can be expressed with the same set of variables presented by equation (6.15)

Note that, except for volatility σ, all variables listed above are known to the investor: option and stock prices can be observed from the existing option contracts and stocks being traded, while exercise price and time to expiration are listed in the option's contract. Therefore, it follows that for each option contract the term σ can be derived. This term represents the future stock volatility that the option price implies, i.e. the *implied volatility*. This framework suggests that if the investor was able to forecast implied volatility better than the market, he could devise an option trading strategy that trades on mispriced options. This study does not attempt to fully simulate such option trading strategy, but it rather tries to assess ASVI's capability to improve implied volatility forecasting models.

6-1 Methodology

Similarly to the approach suggested by (Dimpfl and Jank 2012), the forecasting ability of ASVI is assessed by setting up an autoregressive AR(p) model that will serve as benchmark for the analysis. The model will then be augmented by including a lagged ASVI term. Forecasts produced by benchmark and augmented model will then be compared in- and out-of-sample.

It may be argued that scientific literature has produced models whose forecasting ability is superior to a simple AR(p) model. For example, GARCH specifications take into account the heteroscedasticity that volatility time series usually exhibit. However, it should be remembered that the purpose of this study is not to produce a state of the art forecasting model, but rather to assess ASVI's capability to improve the prediction

An autoregressive model of lag order *p* used as benchmark can be written as

$$y_t = \alpha + \sum_{i=1}^{p} \beta_i y_{t-i} + \varepsilon_t \tag{6.2}$$

Where y_t denotes the series to be forecasted. As equation (6.2) suggests, values of the forecasted series on time *t* are linearly dependent on the previous known realizations on time *t-1*, *t-2,…,t-p*. In this study, lag orders *p=1,2,3* will be considered. The in-sample period used corresponds to about one third of the total sample size and it ranges from 1/1/2007 to 31/12/2009. The out of sample period ranges from 1/1/2010 to 31/11/2015.

Out-of-sample forecasting analysis is generally considered to be more reliable than the in-sample one, because the coefficients of the in-sample fitted model are calculated by using the whole in sample period. In other words, in-sample forecasting procedures "cheat" by looking at periods ahead and deriving optimal coefficients that include information that was not available on time *t.* For this reason, in-sample forecasting is expected to be more accurate (and overly optimistic) than the out-of-sample one. Out-of-sample forecasting provides a more accurate simulation of the model capabilities, because forecasted values at time *t* are obtained by using only information that was available on previous periods. More specifically, the information used to forecast each observation $\widehat{y_t}$ depends on the estimation window used.

In order to produce out-of-sample forecast $\widehat{y_t}$ the model described by equation (6.2) is estimated over period $[t-w-1, t-1]$, where *k* denotes the estimation window used. The fitted model is then used to generate a 1-step-ahead forecast. The estimation window is then rolled ahead by one period in order to compute the next forecast observation; i.e., the model is fit on period $[t-w, t]$ in order to determine $\widehat{y_{t+1}}$.

The aforementioned approach requires the choice of an appropriate window length *w.* There are no hard rules regarding the optimal value for this parameter, however, a few considerations should be made. The choice of window length involves a balance between two opposing factors. a shorter window implies a smaller dataset used to generate the model's coefficients, which may result in overfitting the model to the specific window in use, incorporating effects that are only temporary. On the other hand, a long observation window increases the chances that the data-generating process has mutated over the estimation period covered, resulting in a model that is based on data which is no longer representative of the current behavior. Because of these considerations and in order to test the robustness of the results, different estimation windows will be used: models will be computed for w = *60,150,250,375*. The notation of the models thus becomes $AR(p)^w$.

In addition to the use of different evaluation windows, for each model it is also computed one additional forecast that is given by the average of the forecasts obtained using different window lengths, i.e. the average the forecasts obtained with models $AR(p)^{w=60}$, $AR(p)^{w=150}$, $AR(p)^{w=250}$ and $AR(p)^{w=375}$. Previous scientific literature suggests that averaging forecasts of different models often entail better estimates. Pesaran and Pick (2011) show that averages of models that only differ in window length lead to lower Root Mean Squared Errors. More interestingly, they argue that this approach mitigates the forecasting errors due to structural breaks in the time series.

6-1-2 Forecast Evaluation Criteria

All in- and out-of-sample forecasts are evaluated by a series of indicators in order to test their accuracy and unbiasedness. Suppose the forecast period is $j = T + 1, T + 2, \dots, T + h$ and let y_t and $\widehat{y_t}$ denote the actual and forecasted values respectively at time *t.* The following measures can be computed:

Mean Absolute Error (MAE)

$$MAE = \sum_{t=T+1}^{T+h} \frac{|\widehat{y_t} - y_t|}{h} \tag{6.3}$$

This indicator measures the average error of the forecasted series, since the absolute value of errors is used, errors with opposite sign will not cancel each other out, providing a reliable estimation of forecast's accuracy: the smaller the error, the more accurate the forecast is. MAE places equal weight on all forecast errors.

Mean Absolute percentage error (MAPE)

$$MAPE = 100 \sum_{t=T+1}^{T+h} \frac{\left|\frac{\widehat{y_t} - y_t}{y_t}\right|}{h} \tag{6.4}$$

MAPE measures the mean absolute percentage deviation of the forecasted series. Due to its intuitive interpretation, MAPE is one of the most commonly used forecast accuracy indicators. However, scientific literature as highlighted a series of issues that affect this measure (cit). Most notably, MAPE puts heavier penalty on negative errors.

Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\sum_{t=T+1}^{T+h} \frac{(\widehat{y_t} - y_t)^2}{h}} \tag{6.5}$$

RMSE is another measurement of forecast accuracy. Forecast errors are squared so that opposite signs will not cancel each other out. Because of the squared term, RMSE places higher weight on highest errors compared to MAE

Theil Inequality Coefficient (U)

$$Theil\ U = \frac{\sqrt{\frac{1}{h}\sum_{t=T+1}^{T+h}(\widehat{y_t} - y_t)^2}}{\sqrt{\frac{1}{h}\sum_{t=T+1}^{T+h}\widehat{y_t}^{\,2}} + \sqrt{\frac{1}{h}\sum_{t=T+1}^{T+h}y_t^{\,2}}} \tag{6.6}$$

Theil inequality coefficient is a scale invariant measure: as equation (6.6) indicates, the numerator is constituted by RMSE, while the denominator scales Theil's inequality statistics so that it always lies between one and zero, with zero indicating a perfect fit. As Pindyck and Rubinfeld (1998)) show, it is possible to decompose the mean squared errors (MSE) in equation (6.6) and derive the following measures:

$$Bias\ proportion\ (U_B) = \frac{\left[\left(\frac{1}{h}\sum \widehat{y_t}\right) - \overline{y_t}\right]^2}{\frac{1}{h}\sum(\widehat{y_t} - y_t)^2} = \frac{\left[\left(\frac{1}{h}\sum \widehat{y_t}\right) - \overline{y_t}\right]^2}{MSE} \tag{6.7}$$

$$Variance\ proportion\ (U_v) = \frac{(\sigma_{\widehat{y_t}} - \sigma_{y_t})^2}{\frac{1}{h}\sum(\widehat{y_t} - y_t)^2} = \frac{(\sigma_{\widehat{y_t}} - \sigma_{y_t})^2}{MSE} \tag{6.8}$$

$$Covariance\ proportion\ (U_c) = \frac{2(1-c)\sigma_{\widehat{y_t}}\sigma_{y_t}}{\frac{1}{h}\sum(\widehat{y_t} - y_t)^2} = \frac{2(1-c)\sigma_{\widehat{y_t}}\sigma_{y_t}}{MSE} \tag{6.9}$$

Where σ_{y_t} and $\sigma_{\widehat{y_t}}$ denote the standard deviation of y_t and $\widehat{y_t}$ respectively, and *c* indicates the correlation between the two series. Note that, by construction, these three measures always add up to one.

Bias proportion U_B provides a measure of systematic error, i.e. how far the mean of the forecast is from the mean of the actual. Clearly, regardless of the value of Theil U, a lower bias proportion is preferable a lower bias proportion. Variance proportion U_v measures how far the variation of the forecast is from the variation of the actual series. As argued by Pindyck and Rubinfeld (1998), U_v indicates the ability of the model to replicate the degree of variability of the forecasted series: a high value indicates that the forecasted series fluctuated much more or much less than the actual. This property is clearly undesirable and therefore lower values of U_v indicate a better forecast.

Lastly, covariance proportion measures the remaining unsystematic error. Based on Pindyck and Rubinfeld (1998). considerations, it follows that a high U_c is preferable, as it entails lower U_B and U_v.

Mincer-Zarnowitz $\boldsymbol{R^2}$

Another measure of forecast accuracy is Mincer-Zarnowitz (MZ) R^2, that is derived from the following OLS regression:

$$y_t = b_0 + b_1\widehat{y_t} + \varepsilon_t \tag{6.10}$$

Therefore, MZ R^2 can be written as:

$$MZ\ R^2 = 1 - \frac{\sum_{t=1}^{T}(y_t - \widehat{y_t})^2}{\sum_{t=1}^{T}(y_t - \overline{y_t})^2} \tag{6.11}$$

Where a high R^2 indicates a good fit of the model used.

Comparing forecast accuracy: test of significance

As previously mentioned, when comparing two or more forecast models a reduction in MAE suggests a better performance and a more accurate forecast. However, it is also necessary to test whether such reduction is statistically significant. The intuition behind this test is as follows: let y_t denote the actual observations to be forecasted, while $\widehat{y_{t,1}}$ and $\widehat{y_{t,2}}$ indicate the two competing.

Forecast errors can be written as:

$$e_{t,1} = \widehat{y_{t,1}} - y_t \tag{6.12}$$

$$e_{t,2} = \widehat{y_{t,2}} - y_t \tag{6.13}$$

And the loss function of the forecasts is given by[14]:

$$g\left(y_t, \widehat{y_{t,1}}\right) = \left|\widehat{y_{t,1}} - y_t\right| = \left|e_{t,1}\right| \tag{6.14}$$

$$g\left(y_t, \widehat{y_{t,2}}\right) = \left|\widehat{y_{t,2}} - y_t\right| = \left|e_{t,2}\right| \tag{6.15}$$

Thus, it follows that the loss differential between the two forecasts is:

$$d_t = g\left(y_t, \widehat{y_{t,1}}\right) - g\left(y_t, \widehat{y_{t,2}}\right) = \left|e_{t,1}\right| - \left|e_{t,2}\right| \tag{6.16}$$

And the two forecasts have the same accuracy if and only if the loss differential has zero expectation for all *t*. Therefore, the null hypothesis will be:

$$H_0 : E(d_t) = 0 \; for \; all \; t \tag{6.17}$$

Versus the alternative hypothesis:

$$H_1 : E(d_t) = \mu \neq 0 \tag{6.18}$$

A standard t-test is run on H_0 , where a largely positive t-statistics indicates that $\widehat{y_2}$ is more accurate than $\widehat{y_1}$. Conversely, a largely negative t-stat suggests that $\widehat{y_1}$ is more accurate than $\widehat{y_2}$. This procedure is consistent with the approach suggested by (Diebold 2015) and (Harvey, Leybourne, and Newbold 1997), and it only differs in the way standard errors are computed.

[14] Instead of absolute errors, squared errors can also be used. In this study, absolute errors will be preferred as they better suit the normality assumptions on which the test is based

6-2 Results

Table 6-1 reports the in-sample forecast evaluation of AR(p) models and their augmented specifications $AR(p) + \ topic\ ASVI^{k=40}$. Models are estimated over the full in-sample period that ranges from 1/1/2007 to 31/12/2009.

For each model pair (AR; AR+ASVI), the best performing indicator is marked bold in order to facilitate the interpretation. Regardless of the lag specification p used, augmented model specifications seem to produce the best results. RMSE, MAE and MAPE report that smaller errors are achieved when ASVI is included in the model specification. As the t-stat highlights, MAE reduction is statistically significant for all lag orders selected. However, bias proportion of Theil inequality is sometimes higher. Overall, topic ASVI seems to incorporate information that can improve the quality of implied volatility forecasts.

Augmented models have also been computed using different topic ASVI indicators (e.g. $topic\ ASVI^{k=20}$) , leading to results consistent to the ones presented above. Therefore, these results are not reported for sake of brevity

Table 6-1: In-sample results for $AR(p)$ model and its augmented specification $AR(p) + \ topic\ ASVI^{k=40}$

Model	$AR(1)$	$AR(1) + ASVI$	$AR(2)$	$AR(2) + ASVI$	$AR(3)$	$AR(3) + ASVI$
RMSE	0,057635	**0,057513**	0,056925	**0,056843**	0,056756	**0,056699**
MAE	0,037909	**0,037883**	0,03785	**0,037817**	0,037813	**0,037785**
t-stat		(-2,75)		(-7,92)		(-7,13)
MAPE	1,031569	**1,030818**	1,029249	**1,028247**	1,028047	**1,027191**
Theil U	0,007792	**0,007774**	0,007693	**0,00768**	0,007667	**0,007658**
Bias	**0,012696**	0,012764	0,011627	**0,011427**	**0,011172**	0,0112
Variance	0,004208	**0,003887**	0,00548	**0,004946**	0,006677	**0,006234**
Covariance	0,983096	**0,983349**	0,982893	**0,983627**	0,982151	**0,982566**
MZ R^2	0,98405	**0,984122**	0,98441	**0,98446**	0,984477	**0,984513**

Table 6-2 reports the in-sample forecast evaluation of AR(p) models and their augmented specifications $AR(p) + \ ticker\ ASVI^{k=40}$. It is important to note that un-augmented models AR(p) have the same specifications of the ones presented in the previous table 6-1. However, in order to ensure comparability with their augmented versions, forecast analysis is conducted on a restricted sample that only includes those firms where a valid ASVI observation is available. In other words, AR(p) specifications of table 6-1 are evaluated only on those 411 firms with valid topic SVI, while AR(p) specifications of table 6-2 are evaluated only on the 122 firms with valid ticker SVI.

All forecast analysis indicators seem to suggest that the augmented models outperform the simple AR(p) specifications. This result holds regardless of the lag order *p* selected. Unlike topic SVI, ticker SVI seems to positively reduce the bias proportion of Theil U.

MAE reduction t-statistic is always negative, because the augmented model produces smaller absolute errors. However, unlike the results in table 6-1, the reduction is only significant for *p=2*. This suggests that the accuracy of the augmented model may be only marginally superior to the benchmark, and the results are unlikely to hold out-of-sample.

Table 6-2: In-sample results for $AR(p)$ model and its augmented specification $AR(p) + \ ticker\ ASVI^{k=40}$

Model	$AR(1)$	$AR(1) + ASVI$	$AR(2)$	$AR(2) + ASVI$	$AR(3)$	$AR(3) + ASVI$
RMSE	0,059291	**0,059287**	0,058643	**0,058633**	0,058465	**0,058455**
MAE	0,039045	**0,039044**	0,03895	**0,038939**	0,03891	**0,038908**
t-stat		(-0,43)		(-3,30)		(-0,56)
MAPE	1,053105	**1,053078**	1,049601	**1,049278**	1,048156	**1,048101**
Theil U	0,00792	**0,007919**	0,00783	**0,007828**	0,007802	**0,007801**
Bias prop.	0,016521	**0,016508**	0,014717	**0,014451**	0,014664	**0,014592**
Var prop.	0,004092	**0,004085**	0,005688	**0,005308**	0,00694	**0,006726**
Cov prop.	0,979387	**0,979407**	0,979595	**0,980241**	0,978396	**0,978683**
MZ R^2	0,98533	**0,985332**	0,985619	**0,985624**	0,985681	**0,985686**

Table 6-3 reports the out-of-sample forecast evaluation of AR(p) models and their augmented forms $AR(p) + \ topic\ ASVI^{k=40}$. Forecast is computed for a period ranging from 1/1/2010 to 31/12/2015. As explained in section 6-1, the model is fitted using a rolling observation window *w=60,150,250,375*. For each lag order *p* and observation window *w*, the augmented model forms seem to produce more accurate forecasts. However, this is not always true for what concerns MAE: model specifications with shorter observation windows (*w=60,150)* do not seem register a MAE when augmented. On the other hand, augmenting models with longer observation windows seem to always decrease MAE, and the t-stat suggests that this decrease is mostly significant. It should be remembered that an increase in MAE is not incompatible with a decrease in RMSE; this indicates that augmented models reduce the big forecast errors, but the average error increases.

Regarding Theil coefficients, all augmented models seem to provide a better fit than their un-augmented counterparts. However, they produce mixed results in terms of Bias and Variance proportion. This indicates that although augmented models produce more accurate forecasts, the inclusion of SVI introduce bias,

affecting the mean of the forecast. Moreover, the increase of variance proportion indicates that augmenting the model may reduce its ability to reflect the variability of the actual series.

Overall, it can be said that augmented models with longer window lengths seem to perform better than the simple AR(p) models. Conversely, it is difficult to draw conclusions regarding models with shorter window length, because different indicators produce conflicting results.

The out-of-sample evaluation of models augmented with ticker ASVI is presented by table 6-4. In line with the previous table, augmenting the models results in a RMSE decrease. However, the same cannot be said for MAE: augmented models with a short estimation window result in higher MAE than their un-augmented counterpart. Moreover, the MAE reduction achieved by the remaining augmented models is not statistically significant. Theil coefficient proportions highlight a general increase in Bias for the augmented models. This seem to suggest one more time that ticker ASVI does not perform as well as topic ASVI.

Table 6-5 presents the results of the forecast averages, i.e. the average of out-of-sample forecasts computed using the same model but different estimation windows. Not surprisingly, results are consistent with the individual forecasts analyzed so far: *topic SVI* produces a more significant forecast improvement than *ticker ASVI*. The error reduction is significant for AR(1) and AR(2) in the case of *topic SVI*, while it is never significant for ticker SVI. All indicators except for Theil proportions seem to suggest that topic SVI slightly improve the forecast.

Table 6-3: Out-of-sample results for $AR(p)$ model and its augmented specification $AR(p) + \ topic\ ASVI^{k=40}$. Best performing indicators for each pair are market in bold

Model	$AR(1)^{w=60}$	$AR(1)^{w=60} + ASVI$	$AR(1)^{w=150}$	$AR(1)^{w=150} + ASVI$	$AR(1)^{w=250}$	$AR(1)^{w=250} + ASVI$	$AR(1)^{w=375}$	$AR(1)^{w=375} + ASVI$
RMSE	0,064016	**0,064002**	0,063857	**0,063841**	0,06381	**0,063792**	0,063822	**0,063802**
MAE	0,039069	**0,039065**	0,03884	**0,038832**	0,038765	**0,038753**	0,038739	**0,038725**
t-stat		(-0,22)		(-2,04)		(-3,97)		(-5,69)
MAPE	1,22124	**1,221129**	1,214283	**1,214074**	1,212064	**1,211759**	1,211368	**1,210989**
Theil U	0,009253	**0,009249**	0,00923	**0,009226**	0,009223	**0,009219**	0,009224	**0,00922**
Bias prop.	**0,086793**	0,088045	**0,08798**	0,089246	**0,087721**	0,088908	**0,085505**	0,086659
Var prop.	0,065657	**0,065578**	0,066383	**0,066345**	0,065835	**0,06579**	**0,063526**	0,063575
Cov prop.	**0,84755**	0,846377	**0,845637**	0,844409	**0,846444**	0,845302	**0,850969**	0,849766
MZ R^2	0,967171	**0,967185**	0,967333	**0,967349**	0,967382	**0,9674**	0,967374	**0,967394**

Model	$AR(2)^{w=60}$	$AR(2)^{w=60} + ASVI$	$AR(2)^{w=150}$	$AR(2)^{w=150} + ASVI$	$AR(2)^{w=250}$	$AR(2)^{w=250} + ASVI$	$AR(2)^{w=375}$	$AR(2)^{w=375} + ASVI$
RMSE	0,06253	**0,062511**	0,062318	**0,062296**	0,062239	**0,062215**	0,063857	**0,063841**
MAE	**0,039036**	0,039038	0,038761	**0,038758**	0,038636	**0,038629**	0,03884	**0,038832**
t-stat		(1,95)		(0,21)		(-1,46)		(-2,04)
MAPE	**1,218672**	1,218743	1,210308	**1,210219**	1,206592	**1,206403**	1,214283	**1,214074**
Theil U	0,009034	**0,00903**	0,009004	**0,008999**	0,008992	**0,008988**	0,00923	**0,009226**
Bias prop.	**0,108398**	0,109946	**0,109367**	0,110972	**0,10902**	0,110419	**0,08798**	0,089246
Var prop.	0,067656	**0,067591**	0,069612	**0,069549**	0,069973	**0,069921**	0,066383	**0,066345**
Cov prop.	**0,823946**	0,822462	**0,821021**	0,819478	**0,821007**	0,81966	**0,845637**	0,844409
MZ R^2	0,968655	**0,968674**	0,968864	**0,968886**	0,968943	**0,968966**	0,967333	**0,967349**

Model	$AR(3)^{w=60}$	$AR(3)^{w=60} + ASVI$	$AR(3)^{w=150}$	$AR(3)^{w=150} + ASVI$	$AR(3)^{w=250}$	$AR(3)^{w=250} + ASVI$	$AR(3)^{w=375}$	$AR(3)^{w=375} + ASVI$
RMSE	0,062331	**0,062319**	0,061998	**0,061977**	0,061894	**0,06187**	0,061864	**0,061835**
MAE	**0,039099**	0,039113	**0,03872**	**0,03872**	0,038578	**0,038574**	0,038493	**0,038482**
t-stat		(5,47)		(1,42)		(-0,44)		(-3,19)
MAPE	**1,220001**	1,220431	**1,208512**	1,208527	1,204328	**1,204212**	1,201814	**1,201512**
Theil U	0,009002	**0,008999**	0,008954	**0,00895**	0,008939	**0,008934**	0,008934	**0,008929**
Bias prop.	**0,115344**	0,116913	**0,117482**	0,11922	**0,117088**	0,118591	**0,114975**	0,116467
Var prop.	0,067828	**0,067682**	0,070582	**0,070485**	0,07134	**0,071266**	**0,069785**	0,069832
Cov prop.	**0,816828**	0,815405	**0,811937**	0,810295	**0,811572**	0,810143	**0,815239**	0,813701
MZ R^2	0,968837	**0,968849**	0,969165	**0,969186**	0,969268	**0,969291**	0,969299	**0,969328**

Table 6-4: Out-of-sample results for $AR(p)$ model and its augmented specification augmented specification $AR(p) + \ ticker\ ASVI^{k=40}$.

Model	$AR(1)^{w=60}$	$AR(1)^{w=60} + ASVI$	$AR(1)^{w=150}$	$AR(1)^{w=150} + ASVI$	$AR(1)^{w=250}$	$AR(1)^{w=250} + ASVI$	$AR(1)^{w=375}$	$AR(1)^{w=375} + ASVI$
RMSE	0,064921	**0,06492**	0,064773	**0,064767**	0,064725	**0,064718**	0,064729	**0,064722**
MAE	**0,039575**	0,039582	**0,039368**	**0,039368**	0,0393	**0,039298**	0,039278	**0,039276**
t-stat		(2,53)		(1,65)		(0,00)		(-1,04)
MAPE	**1,224677**	1,224874	**1,218349**	1,218356	1,216367	**1,216302**	1,215771	**1,215706**
Theil U	**0,006901**	**0,006901**	0,006886	**0,006885**	**0,00688**	**0,00688**	0,006881	**0,00688**
Bias prop.	0,086101	**0,086052**	0,085691	**0,085664**	0,086462	**0,086438**	0,08848	**0,088465**
Var prop.	0,077657	**0,077651**	**0,077729**	0,077754	**0,078134**	0,078151	**0,080172**	0,080179
Cov prop.	0,836242	**0,836297**	0,83658	**0,836582**	0,835404	**0,835411**	0,831348	**0,831356**
MZ R^2	0,97193	**0,971931**	0,972057	**0,972063**	0,972099	**0,972105**	0,972099	**0,972105**

Model	$AR(2)^{w=60}$	$AR(2)^{w=60} + ASVI$	$AR(2)^{w=150}$	$AR(2)^{w=150} + ASVI$	$AR(2)^{w=250}$	$AR(2)^{w=250} + ASVI$	$AR(2)^{w=375}$	$AR(2)^{w=375} + ASVI$
RMSE	**0,063011**	**0,063011**	0,062795	**0,062786**	0,062682	**0,062672**	0,062677	**0,062666**
MAE	**0,039511**	0,039524	**0,039254**	0,039255	0,039131	**0,039129**	0,039077	**0,039074**
t-stat		(3,57)		(0,41)		(-0,87)		(-1,36)
MAPE	**1,221539**	1,2219	**1,213663**	1,213685	1,210059	**1,209972**	1,20846	**1,208347**
Theil U	**0,006694**	**0,006694**	0,006671	**0,00667**	0,006659	**0,006658**	0,006658	**0,006657**
Bias prop.	0,075078	**0,075036**	0,075032	**0,075008**	0,076009	**0,076002**	**0,07776**	0,077768
Var prop.	**0,08404**	0,08413	**0,083168**	0,083233	**0,082888**	0,082932	**0,084508**	0,084542
Cov prop.	**0,840881**	0,840834	**0,8418**	0,841759	**0,841103**	0,841065	**0,837732**	0,83769
MZ R^2	**0,973546**	0,973545	0,973725	**0,973732**	0,973818	**0,973827**	0,973824	**0,973834**

Model	$AR(3)^{w=60}$	$AR(3)^{w=60} + ASVI$	$AR(3)^{w=150}$	$AR(3)^{w=150} + ASVI$	$AR(3)^{w=250}$	$AR(3)^{w=250} + ASVI$	$AR(3)^{w=375}$	$AR(3)^{w=375} + ASVI$
RMSE	**0,062743**	0,062749	0,062424	**0,062416**	0,062256	**0,062245**	0,062236	**0,062224**
MAE	**0,039586**	0,039607	**0,039227**	0,03923	0,039076	**0,039075**	0,038993	**0,03899**
t-stat		(5,41)		(1,03)		(-0,37)		(-1,02)
MAPE	**1,223193**	1,223803	**1,212201**	1,212281	1,207774	**1,207727**	1,205281	**1,20519**
Theil U	**0,006661**	0,006662	**0,006627**	**0,006627**	0,006609	**0,006608**	0,006607	**0,006606**
Bias prop.	0,06848	**0,068391**	0,068045	**0,068023**	0,069275	**0,069269**	**0,070863**	0,070882
Var prop.	**0,084989**	0,085055	**0,084096**	0,084169	**0,083654**	0,083702	**0,085093**	0,085138
Cov prop.	0,846531	**0,846554**	**0,847858**	0,847808	**0,847071**	0,847029	**0,844044**	0,843979
MZ R^2	**0,973759**	0,973754	0,974023	**0,974029**	0,974161	**0,97417**	0,974179	**0,974189**

Table 6-5: Average of forecasts models computed over different estimation windows.

Model:	$AR(1)$	$AR(1) + topic\ SVI$	$AR(2)$	$AR(2) + topic\ SVI$	$AR(3)$	$AR(3) + topic\ SVI$
RMSE	0,063851	**0,063833**	0,06228	**0,062254**	0,061953	**0,061928**
MAE	0,038824	**0,038813**	0,038703	**0,038694**	0,038661	**0,038656**
		(-3,61)		(-1,90)		(-0,38)
MAPE	1,213823	**1,213527**	1,20857	**1,208339**	1,206764	**1,206643**
Theil U	0,009229	**0,009225**	0,008998	**0,008993**	0,008947	**0,008942**
Bias prop.	**0,087066**	0,088285	**0,108567**	0,110059	**0,116476**	0,118068
Var prop.	0,065435	**0,065411**	0,069027	**0,069008**	0,070134	**0,070079**
Cov prop.	**0,847499**	0,846304	**0,822406**	0,820933	**0,81339**	0,811853
MZ R^2	0,96734	**0,967359**	0,968903	**0,968929**	0,96921	**0,969235**

Model:	$AR(1)$	$AR(1) + ticker\ SVI$	$AR(2)$	$AR(2) + ticker\ SVI$	$AR(3)$	$AR(3) + ticker\ SVI$
RMSE	0,064765	**0,064759**	0,062738	**0,062729**	0,062338	**0,06233**
MAE	**0,039354**	**0,039354**	**0,039196**	**0,039196**	**0,039157**	0,039159
		(-0,16)		(-0,09)		(0,72)
MAPE	1,217976	**1,217957**	1,211985	**1,211962**	**1,210164**	1,210211
Theil U	0,006885	**0,006884**	0,006665	**0,006664**	0,006618	**0,006617**
Bias prop.	0,08674	**0,086713**	0,076093	**0,07608**	0,069329	**0,069309**
Var prop.	**0,078441**	0,078452	**0,083711**	0,083771	**0,084547**	0,084608
Cov prop.	0,83482	**0,834835**	**0,840197**	0,840149	**0,846124**	0,846083
MZ R^2	0,972065	**0,97207**	0,973772	**0,97378**	0,974094	**0,974101**

Chapter 7

Conclusion

Based on the insights that online search query volumes constitute a measure of revealed attention and they can serve as a proxy for investor sentiment, it was suggested that SVI embeds information which explains and anticipates financial markets' dynamics. Daily SVI observations were collected for all firms that composed the S&P500 index from January 2007 to November 2015. More specifically, SVI was downloaded in two variants (*topic SVI* and *ticker SVI)* in order to determine which type of search queries are better suited to capture investor attention. SVI reports obtained were merged and manipulated in order to compute an indicator of abnormal search volume (ASVI).

Subsequently, three studies were conducted in order to determine ASVIs capability to forecast important financial indicators. The first study (Chapter 4) focused on the relation between ASVI and trading activity. A positive correlation between abnormal stock turnover and both ASVI series (*topic SVI in particular*) was proven to exist. More interestingly, when computing time lagged cross correlations, positive lag orders seemed to produce higher estimates than the negative ones, suggesting that the causal relation "ASVI anticipates trading activity" was stronger than "trading activity anticipates ASVI". This intuition was corroborated by a series of Granger-causality tests.

Second study (Chapter 5) was based on the intuition that an increase in attention (SVI increase) indicates investor's interest towards a specific firm, which can translate in their decision to buy stocks driving the price upward. Therefore, in accordance to Fama-Macbeth approach, a series of rolling regressions were conducted in order to produce a preliminary analysis that could indicate the relationship between ASVI and different types of returns. The most suitable ASVI indicator (*topic SVI)* was selected and used in order to compute a series of long-short trading strategies based on SVI-sorted portfolios. The analysis of these trading strategies highlighted that as the time interval between SVI observation and trading activity increased, abnormal returns decreased. This provides evidence that although ASVI embeds information that anticipates high returns, exploiting such phenomena is very difficult: markets align to information rapidly, and transaction costs hinder trading profitability even further.

Lastly, Chapter 6 assessed whether SVI carries predictive information pertaining stock implied volatility. Several specifications of an AR(p) model were introduced with the purpose to provide a benchmark to the

analysis. Autoregressive models were then augmented with the inclusion of SVI terms. The forecasting power of these models was assessed both in-sample and out-of-sample. It was shown that those models who best performed in-sample (topic SVI models) could outperform benchmark models out-of-sample, provided that an appropriate estimation window was used.

Chapter 8

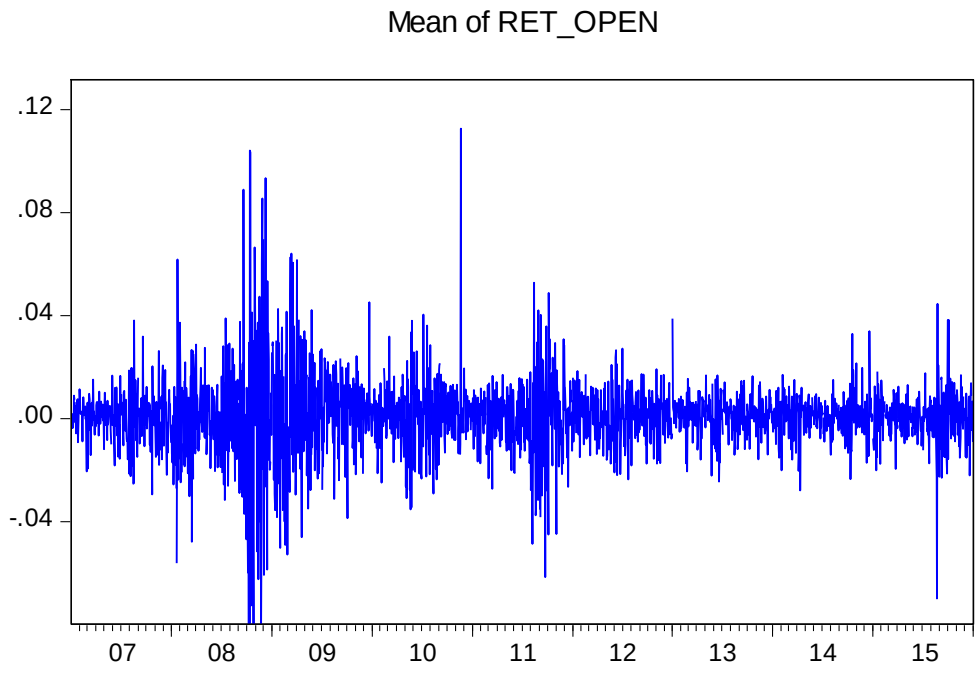
Future Research

There is no doubt that SVI data poses itself to several fascinating applications. However, this study has also highlighted that several limitations must be faced by the researcher: the lack of an API interface complicates data mining, and data availability is often uncertain. However, it should be noted that the situation is constantly improving, paving the way to new approaches that were not possible until just a few months ago. As an example, Google Trends made available, starting from 2015, SVI data in real time. As it has been shown by the study conducted on stock returns, using a timely indicator is essential in order to "beat the market", and the use of real time data could provide encouraging results.

In addition, more elaborated approaches could be used to select the queries to be used for the research. A possible approach could be to designate an in-sample period to be used in order to determine which queries are the most correlated with financial indicators and then to trade only stocks associated with the selected queries in a subsequent out-of-sample period.

Appendix

A1: Cross sectional mean values of open-to-open returns over time. Some sudden spikes suggest the existence of outliers. Although this issue does not affect many observations, the error size is very big and had to be corrected in order to ensure reliable results.



A2: Correlation of selected turnover variables with topic SVI computed over different observation windows

Turnover variable:	ATURN5										
	lag order(δ)										
	-5	-4	-3	-2	-1	0	1	2	3	4	5
LOG(NORM_TOPIC_SVI)	0,24%	0,62%	0,63%	0,83%	1,42%	2,44%	1,13%	-0,07%	-0,67%	-0,89%	-0,86%
TOPIC_SVI5	-3,64%	-2,41%	-1,84%	0,56%	5,48%	12,60%	7,75%	2,45%	-0,66%	-2,22%	-2,64%
TOPIC_SVI10	-0,66%	1,72%	2,43%	3,85%	7,19%	**12,63%**	6,83%	1,24%	-1,70%	-3,00%	-3,08%
TOPIC_SVI20	0,97%	2,78%	2,99%	3,95%	6,76%	11,58%	5,80%	0,40%	-2,34%	-3,46%	-3,42%
TOPIC_SVI40	1,02%	2,63%	2,77%	3,61%	6,13%	10,48%	5,09%	0,10%	-2,42%	-3,41%	-3,34%
TOPIC_SVI60	1,03%	2,52%	2,62%	3,38%	5,74%	9,81%	4,71%	0,01%	-2,36%	-3,29%	-3,20%

turnover variable:	ATURN10										
	lag order(δ)										
	-5	-4	-3	-2	-1	0	1	2	3	4	5
LOG(NORM_TOPIC_SVI)	0,33%	0,70%	0,74%	1,02%	1,71%	2,86%	1,79%	0,80%	0,30%	-0,11%	-0,42%
TOPIC_SVI5	-4,15%	-3,45%	-3,38%	-1,26%	3,68%	11,43%	8,28%	4,94%	3,45%	2,01%	0,67%
TOPIC_SVI10	-2,19%	0,15%	1,12%	3,17%	7,34%	**13,87%**	9,60%	5,39%	3,30%	1,42%	-0,12%
TOPIC_SVI20	1,00%	2,84%	3,24%	4,62%	7,93%	13,47%	8,84%	4,47%	2,29%	0,41%	-1,06%
TOPIC_SVI40	1,30%	2,88%	3,14%	4,32%	7,26%	12,22%	7,84%	3,76%	1,72%	-0,01%	-1,36%
TOPIC_SVI60	1,35%	2,81%	3,02%	4,09%	6,83%	11,46%	7,30%	3,44%	1,50%	-0,14%	-1,39%

turnover variable:	ATURN40										
	lag order(δ)										
	-5	-4	-3	-2	-1	0	1	2	3	4	5
LOG(NORM_TOPIC_SVI)	0,75%	1,11%	1,16%	1,45%	2,17%	3,38%	2,50%	1,69%	1,33%	1,05%	0,85%
TOPIC_SVI5	-4,31%	-3,89%	-4,16%	-2,49%	2,00%	9,38%	6,67%	3,83%	2,87%	2,13%	1,56%
TOPIC_SVI10	-3,44%	-1,50%	-0,83%	0,97%	5,04%	11,65%	8,33%	5,12%	3,97%	3,08%	2,47%
TOPIC_SVI20	-0,28%	1,59%	2,13%	3,70%	7,29%	13,20%	9,65%	6,29%	4,98%	3,95%	3,23%
TOPIC_SVI40	1,87%	3,49%	3,88%	5,19%	8,34%	**13,60%**	10,12%	6,87%	5,51%	4,44%	3,65%
TOPIC_SVI60	2,41%	3,87%	4,17%	5,34%	8,23%	13,10%	9,76%	6,64%	5,31%	4,25%	3,48%

turnover variable:	ATURN60										
	lag order(δ)										
	-5	-4	-3	-2	-1	0	1	2	3	4	5
LOG(NORM_TOPIC_SVI)	0,87%	1,23%	1,29%	1,59%	2,31%	3,52%	2,67%	1,89%	1,56%	1,29%	1,11%
TOPIC_SVI5	-4,43%	-4,03%	-4,32%	-2,68%	1,76%	9,04%	6,42%	3,65%	2,72%	2,01%	1,47%
TOPIC_SVI10	-3,73%	-1,84%	-1,18%	0,60%	4,64%	11,18%	7,97%	4,85%	3,74%	2,90%	2,34%
TOPIC_SVI20	-0,81%	1,05%	1,60%	3,17%	6,76%	12,64%	9,22%	5,97%	4,73%	3,77%	3,12%
TOPIC_SVI40	1,34%	3,00%	3,43%	4,79%	7,98%	**13,25%**	9,94%	6,83%	5,58%	4,61%	3,93%
TOPIC_SVI60	2,27%	3,79%	4,15%	5,38%	8,33%	13,22%	10,04%	7,05%	5,82%	4,86%	4,17%

turnover variable:	log(turnover)										
	lag order(δ)										
	-5	-4	-3	-2	-1	0	1	2	3	4	5
LOG(NORM_TOPIC_SVI)	-0,75%	-0,52%	-0,46%	-0,27%	0,19%	0,96%	0,47%	0,47%	-0,18%	-0,32%	-0,41%
TOPIC_SVI5	-2,69%	-2,47%	-2,68%	-1,71%	0,99%	5,50%	3,92%	2,24%	1,69%	1,26%	0,95%
TOPIC_SVI10	-2,43%	-1,29%	-0,91%	0,16%	2,64%	6,70%	4,77%	2,88%	2,23%	1,73%	1,41%
TOPIC_SVI20	-0,87%	0,25%	0,58%	1,54%	3,75%	7,42%	5,37%	3,41%	2,68%	2,12%	1,74%
TOPIC_SVI40	0,24%	1,25%	1,53%	2,37%	4,35%	7,65%	5,67%	3,81%	3,08%	2,52%	2,14%
TOPIC_SVI60	0,83%	1,77%	2,01%	2,78%	4,62%	**7,70%**	5,82%	4,04%	3,34%	2,79%	2,42%

A3: Correlation of selected turnover variables with ticker SVI computed over different observation windows

turnover variable:	ATURN5										
	lag order(δ)										
	-5	-4	-3	-2	-1	0	1	2	3	4	5
LOG(NORM_TICKER_SVI)	0,05%	0,16%	0,15%	0,26%	0,53%	0,93%	0,20%	-0,38%	-0,62%	-0,72%	-0,66%
TICKER_SVI5	-1,84%	-1,22%	-0,70%	1,29%	4,79%	**9,23%**	4,94%	1,12%	-0,71%	-1,70%	-1,71%
TICKER_SVI10	0,44%	1,77%	2,25%	3,42%	5,69%	8,73%	4,17%	0,39%	-1,26%	-2,05%	-1,90%
TICKER_SVI20	1,60%	2,48%	2,56%	3,35%	5,12%	7,58%	3,27%	-0,23%	-1,72%	-2,38%	-2,09%
TICKER_SVI40	1,53%	2,24%	2,21%	2,82%	4,36%	6,53%	2,65%	-0,46%	-1,75%	-2,32%	-2,08%
TICKER_SVI60	1,40%	2,03%	1,99%	2,54%	3,95%	5,96%	2,29%	-0,63%	-1,85%	-2,37%	-2,14%

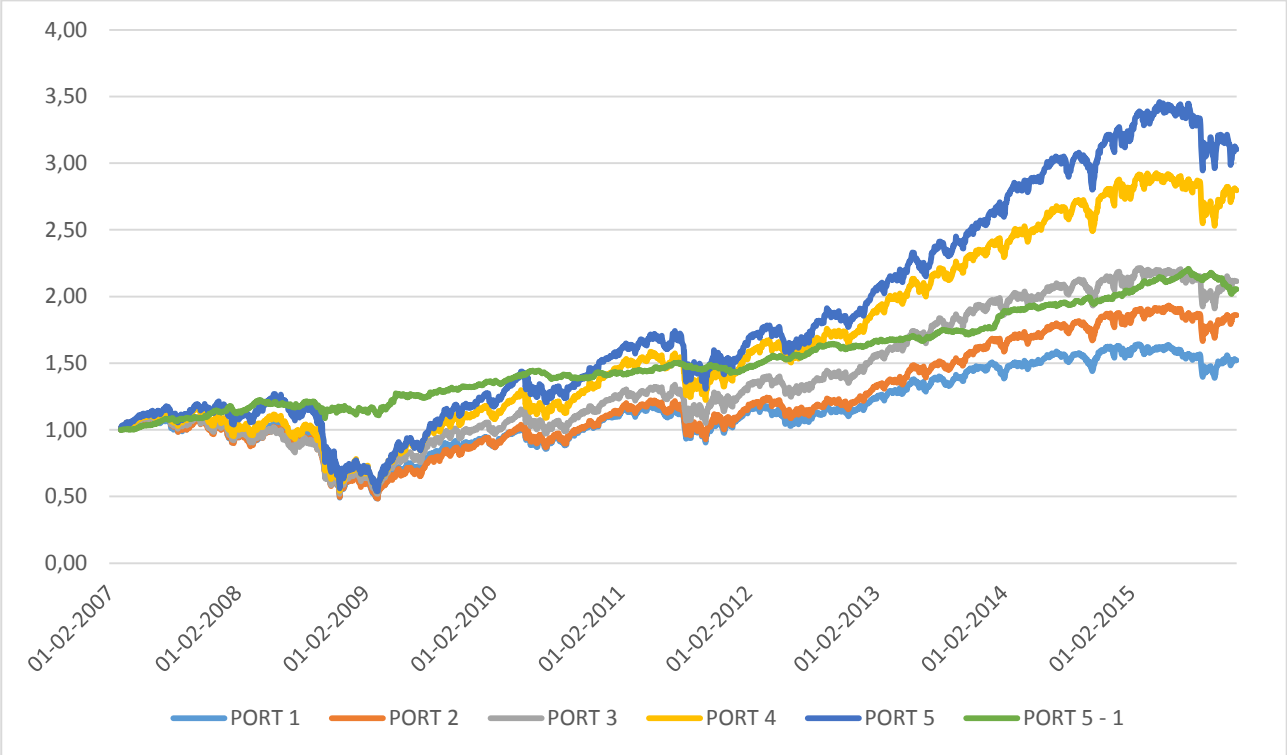
turnover variable:	ATURN10										
	lag order(δ)										
	-5	-4	-3	-2	-1	0	1	2	3	4	5
LOG(NORM_TICKER_SVI)	0,22%	0,35%	0,37%	0,51%	0,83%	1,30%	0,67%	0,17%	-0,07%	-0,29%	-0,42%
TICKER_SVI5	-2,41%	-1,98%	-1,64%	0,19%	3,87%	8,92%	5,95%	3,46%	2,49%	1,40%	0,60%
TICKER_SVI10	-0,58%	0,91%	1,75%	3,38%	6,36%	**10,25%**	6,75%	3,80%	2,45%	1,10%	0,16%
TICKER_SVI20	1,83%	2,92%	3,29%	4,37%	6,57%	9,56%	5,98%	3,02%	1,60%	0,28%	-0,54%
TICKER_SVI40	2,10%	2,93%	3,10%	3,89%	5,74%	8,30%	5,01%	2,32%	1,05%	-0,13%	-0,89%
TICKER_SVI60	1,91%	2,65%	2,78%	3,50%	5,18%	7,56%	4,42%	1,87%	0,66%	-0,46%	-1,17%

turnover variable:	ATURN40										
	lag order(δ)										
	-5	-4	-3	-2	-1	0	1	2	3	4	5
LOG(NORM_TICKER_SVI)	0,35%	0,50%	0,54%	0,71%	1,07%	1,60%	1,08%	0,67%	0,51%	0,36%	0,26%
TICKER_SVI5	-2,79%	-2,55%	-2,45%	-0,85%	2,57%	7,49%	5,00%	2,96%	2,47%	2,00%	1,76%
TICKER_SVI10	-1,71%	-0,41%	0,30%	1,89%	4,91%	9,05%	6,40%	4,26%	3,66%	3,09%	2,79%
TICKER_SVI20	1,00%	2,23%	2,82%	4,18%	6,71%	10,15%	7,47%	5,28%	4,53%	3,84%	3,48%
TICKER_SVI40	3,20%	4,19%	4,55%	5,58%	7,67%	**10,58%**	7,99%	5,88%	5,07%	4,32%	3,83%
TICKER_SVI60	3,44%	4,28%	4,53%	5,40%	7,26%	9,89%	7,35%	5,29%	4,45%	3,68%	3,16%

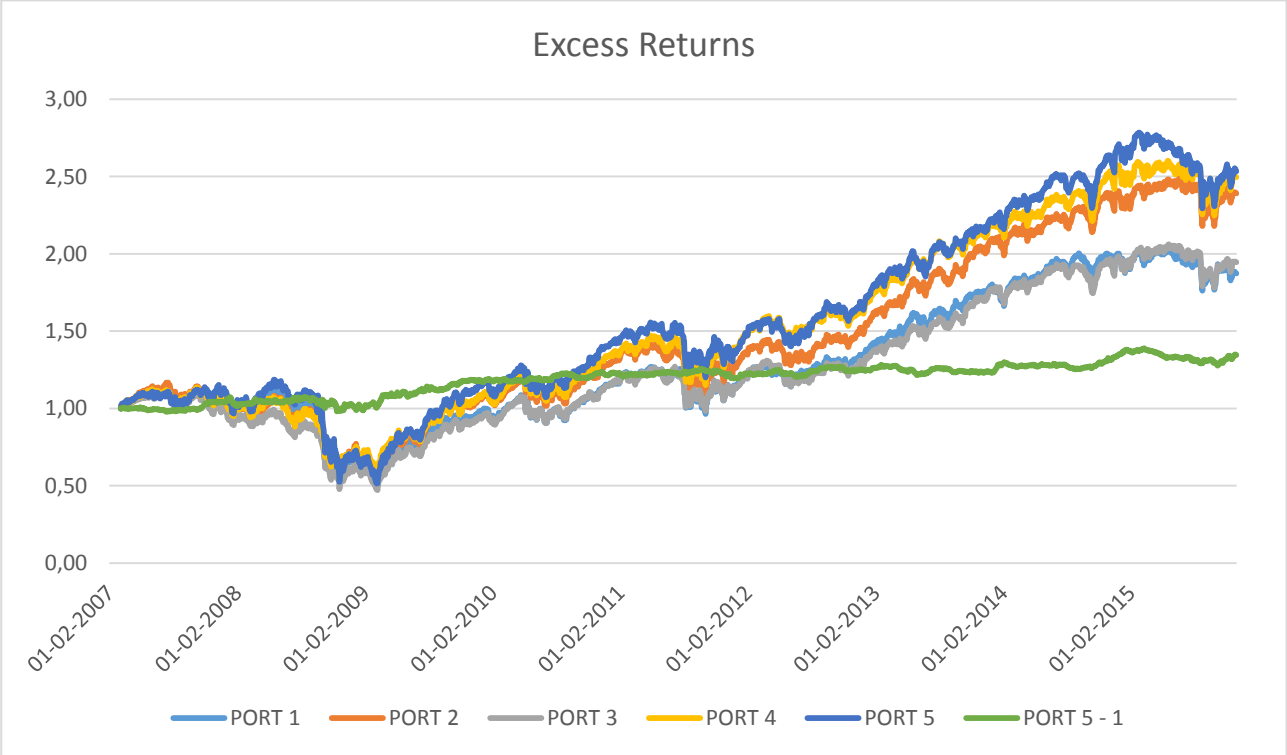
turnover variable:	ATURN60										
	lag order(δ)										
	-5	-4	-3	-2	-1	0	1	2	3	4	5
LOG(NORM_TICKER_SVI)	0,29%	0,45%	0,50%	0,68%	1,04%	1,57%	1,08%	0,68%	0,53%	0,39%	0,31%
TICKER_SVI5	-2,98%	-2,76%	-2,68%	-1,11%	2,27%	7,14%	4,72%	2,72%	2,23%	1,80%	1,59%
TICKER_SVI10	-2,07%	-0,81%	-0,11%	1,47%	4,46%	8,57%	6,00%	3,91%	3,33%	2,82%	2,56%
TICKER_SVI20	0,41%	1,62%	2,21%	3,59%	6,12%	9,58%	7,00%	4,90%	4,19%	3,58%	3,30%
TICKER_SVI40	2,63%	3,66%	4,08%	5,17%	7,31%	**10,28%**	7,83%	5,84%	5,12%	4,49%	4,11%
TICKER_SVI60	3,32%	4,21%	4,53%	5,48%	7,39%	10,08%	7,68%	5,73%	4,97%	4,30%	3,87%

turnover variable:	log(turnover)										
	lag order(δ)										
	-5	-4	-3	-2	-1	0	1	2	3	4	5
LOG(NORM_TICKER_SVI)	-5,71%	-5,61%	-5,57%	-5,46%	-5,22%	-4,88%	-5,18%	-5,41%	-5,49%	-5,57%	-5,62%
TICKER_SVI5	-2,32%	-2,20%	-2,18%	-1,23%	0,85%	3,87%	2,40%	1,19%	0,91%	0,65%	0,53%
TICKER_SVI10	-1,87%	-1,10%	-0,69%	0,28%	2,13%	4,70%	3,14%	1,88%	1,56%	1,25%	1,11%
TICKER_SVI20	-0,49%	0,26%	0,62%	1,47%	3,05%	5,21%	3,66%	2,40%	2,00%	1,64%	1,48%
TICKER_SVI40	0,69%	1,33%	1,60%	2,29%	3,64%	**5,51%**	4,05%	2,87%	2,47%	2,11%	1,90%
TICKER_SVI60	1,06%	1,62%	1,84%	2,45%	3,68%	5,39%	3,97%	2,82%	2,41%	2,04%	1,81%

A3: Cumulative excess returns of Trading Strategy I – "Cheating"



A4: Cumulative excess returns of Trading Strategy II – "Semi-Biased"



Bibliography

Askitas, Nikos, and Klaus F. Zimmermann. "Google econometrics and unemployment forecasting." *German Council for Social and Economic Data (RatSWD) Research Notes* 41 (2009).

Bank, Matthias, Martin Larch, and Georg Peter. "Google search volume and its influence on liquidity and returns of German stocks." *Financial markets and portfolio management* 25.3 (2011): 239-264.

Baeza-Yates, Ricardo, Georges Dupret, and Javier Velasco. "A study of mobile search queries in japan." *Proceedings of the International World Wide Web Conference*. 2007.

Bangwayo-Skeete, Prosper F., and Ryan W. Skeete. "Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach." *Tourism Management* 46 (2015): 454-464.

Barber, Brad M., and Terrance Odean. "All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors." *Review of Financial Studies* 21.2 (2008): 785-818.

Bordino, Ilaria, et al. "Web search queries can predict stock market volumes." *PloS one* 7.7 (2012): e40014.

Carhart, Mark M. "On persistence in mutual fund performance." *The Journal of finance* 52.1 (1997): 57-82.

Choi, Hyunyoung, and Hal Varian. "Predicting initial claims for unemployment benefits." (2009).

Cooper, Crystale Purvis, et al. "Cancer Internet search activity on a major search engine, United States 2001-2003." *Journal of medical Internet research* 7.3 (2005): e36.

Corley, Courtney, et al. "Monitoring Influenza Trends through Mining Social Media." *BIOCOMP*. 2009.

D'Amuri, Francesco. *Predicting unemployment in short samples with internet job search query data*. University Library of Munich, Germany, 2009.

d'Amuri, Francesco, and Juri Marcucci. "The predictive power of Google searches in forecasting unemployment." *Bank of Italy Temi di Discussione (Working Paper) No* 891 (2012).

Da, Zhi, Joseph Engelberg, and Pengjie Gao. "In search of attention." *The Journal of Finance* 66.5 (2011): 1461-1499.

Da, Zhi, Joseph Engelberg, and Pengjie Gao. *In search of earnings predictability*. Working paper, 2010.

Dergiades, Theologos, Costas Milas, and Theodore Panagiotidis. "Tweets, Google trends, and sovereign

spreads in the GIIPS." *Oxford Economic Papers* (2014): gpu046.

Diebold, Francis X. "Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests."*Journal of Business & Economic Statistics* 33.1 (2015): 1-1.

Dimpfl, Thomas, and Stephan Jank. "Can internet search queries help to predict stock market volatility?." *European Financial Management* (2015).

Ettredge, Michael, John Gerdes, and Gilbert Karuga. "Using web-based search data to predict macroeconomic statistics." *Communications of the ACM* 48.11 (2005): 87-92.

Fama, Eugene F., and James D. MacBeth. "Risk, return, and equilibrium: Empirical tests." *The Journal of Political Economy* (1973): 607-636.

Ginsberg, Jeremy, et al. "Detecting influenza epidemics using search engine query data." *Nature* 457.7232 (2009): 1012-1014.

Guo, Shesen, Ganzhou Zhang, and Run Zhai. "A potential way of enquiry into human curiosity." *British Journal of Educational Technology* 41.3 (2010): E48-E52.

Guzman, Giselle. "Internet search behavior as an economic forecasting tool: The case of inflation expectations." *Journal of economic and social measurement* 36.3 (2011): 119-167.

Hand, Chris, and Guy Judge. "Searching for the picture: forecasting UK cinema admissions using Google Trends data." *Applied Economics Letters*19.11 (2012): 1051-1055.

Harvey, David, Stephen Leybourne, and Paul Newbold. "Testing the equality of prediction mean squared errors." *International Journal of forecasting* 13.2 (1997): 281-291.

Ivković, Zoran, and Scott Weisbenner. "Local does as local is: Information content of the geography of individual investors' common stock investments." *The Journal of Finance* 60.1 (2005): 267-306.

Joseph, Kissan, M. Babajide Wintoki, and Zelin Zhang. "Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search." *International Journal of Forecasting* 27.4 (2011): 1116-1127.

McLaren, Nick, and Rachana Shanbhogue. "Using internet search data as economic indicators." *Bank of England Quarterly Bulletin* 2011 (2011): Q2.

Merton, Robert C. "A simple model of capital market equilibrium with incomplete information." *The journal of finance* 42.3 (1987): 483-510.

Pedersen, Lasse Heje. *Efficiently inefficient: how smart money invests and market prices are determined*. Princeton University Press, 2015.

Perrin, Andrew, and Maeve Duggan. 2015. “Americans’ Internet Access: 2000-2015.” *Pew Research Center* (June): 1–13

Pesaran, M. Hashem, and Andreas Pick. "Forecast combination across estimation windows." *Journal of Business & Economic Statistics* (2012).

Pindyck Robert, S., and Daniel L. Rubinfeld. "Econometric models and econometric forecasts." (1998).

Polgreen, Philip M., et al. "Using internet searches for influenza surveillance." *Clinical infectious diseases* 47.11 (2008): 1443-1448.

Ramos, Sofía B., Helena Veiga, and Pedro Latoeiro. *Predictability of stock market activity using Google search queries*. No. ws130605. 2013.

Rose, Daniel E., and Danny Levinson. "Understanding user goals in web search." *Proceedings of the 13th international conference on World Wide Web*. ACM, 2004.

Schmeling, Maik. "Investor sentiment, herd-like behavior and stock returns: Empirical evidence from 18 industrialized countries." *EFMA 2007 Meetings Paper*. 2007.

Suchoy, Tanya. *Query indices and a 2008 downturn: Israeli data*. Bank of Israel. Research Department, 2009.

Vlastakis, Nikolaos, and Raphael N. Markellos. "Information demand and stock market volatility." *Journal of Banking & Finance* 36.6 (2012): 1808-1821.

Yi, Jeonghee, Farzin Maghoul, and Jan Pedersen. "Deciphering mobile search patterns: a study of yahoo! mobile search queries." *Proceedings of the 17th international conference on World Wide Web*. ACM, 2008.