

CORVINUS UNIVERSITY OF BUDAPEST

FINANCE MSc
INVESTMENT BANKING SPECIALIZATION

MSc THESIS

Stock market prediction using Google Trends data

Author:

Kristóf Attila JUHÁSZ

Supervisor:

Milán Csaba BADICS



November 22, 2020

Abstract

Since the 2000s sentiment analysis is getting a more and more popular field of research in many areas. My paper examines whether measuring investor attention and sentiment has a place amongst the investment decision-supporting tools, therefore I conducted a case study on how effectively Google search volume could be utilized for stock market prediction. After a thorough mapping and analysis of the possibly untouched limitations and overlooked biases involved in this field of study, I developed my final model step-by-step based on the article of Preis et al. (2013), Da et al. (2015), and many other acknowledged studies. The outcomes of the model development phases look highly promising, the final model outperformed the benchmark buy-and-hold portfolio to a great extent. Additionally, one of the most substantial contributions of my work was introducing the standardization methods of the search volume datasets which eliminate any look-ahead biases. To my knowledge, my research was also the first one that defined and distinguished between investor attention and sentiment, including the different approaches with which these could be captured and utilized. Furthermore, my paper gives a comprehensive and well-documented guideline for future researchers, enumerating the key points that are fundamental for developing a search volume prediction model.

Acknowledgement

Here I would like to thank the guidance of my supervisor Milán Csaba Badics, who enthusiastically helped me with his professional and academic knowledge. I would like to express my great appreciation for his valuable and constructive suggestions during the planning and development of this research work. His willingness to give his time so generously has been very much appreciated.

Declarations

Budapesti Corvinus Egyetem



TÉMAVEZETŐI NYILATKOZAT

Badics Milán Csaba

(témavezető neve) konzulens kijelentem, hogy

Juhász Kristóf Attila

(hallgató neve), FM9AGF (Neptun kódja)

Stock market prediction using Google Trends data

című szakdolgozata (mesterképzésben diplomadolgozata) benyújtásra alkalmas és védésre ajánlom.

Kelt: Budapest, 2020.11.10.

Badics Milán Csaba
(témavezető neve) konzulens aláírása

2. számú melléklet

Nyilatkozat saját munkáról

Név: JUHÁSZ KRISTÓF ATTILAE-mail cím: kristof.juhasz2@stud.uni-corvinus.huNEPTUN-kód: FH9AG7

A szakdolgozat címe magyarul:

Tőzsdei előrejelzés Google Trends adatok
felhasználásával

A szakdolgozat címe angolul:

Stock market prediction using Google Trends dataSzakszemináriumvezető (vagy konzulens) neve: BADICS MILÁN CSABA

Alulírott JUHÁSZ KRISTÓF ATTILA (hallgató) igazolom, hogy a szakdolgozatom saját munka eredménye. Bizonyos gondolatok, érvek, logikai és matematikai összefüggések más tanulmányokból való átvétele során a hivatkozásra vonatkozó szabályokat teljes mértékben betartottam.


hallgató aláírása

3. számú melléklet

NYILATKOZAT

Név (nyomtatott betűvel): JUHÁSZ KRISTÓF ATTILA

Alapszak

Mesterszak

Egyéb képzési forma

Dolgozatom elektronikus változatának (pdf dokumentum, a megtekintés, a mentés és a nyomtatás engedélyezett, szerkesztés nem) nyilvánosságáról az alábbi lehetőségek közül kiválasztott hozzáférési szabályzat szerint rendelkezem:

☒ **TELJES NYILVÁNOSSÁGGAL**

A könyvtári honlapon keresztül elérhető a Szakdolgozatok/TDK adatbázisban (<http://szd.lib.uni-corvinus.hu/>), a világháló bármely pontjáról hozzáférhető, fentebb jellemzett pdf dokumentum formájában.

☐ **KORLÁTOZOTT NYILVÁNOSSÁGGAL**

A könyvtári honlapon keresztül elérhető a Szakdolgozatok/TDK adatbázisban (<http://szd.lib.uni-corvinus.hu/>), a kizárólag a Budapesti Corvinus Egyetem területéről hozzáférhető, fentebb jellemzett pdf dokumentum formájában.

☐ **NEM NYILVÁNOS**

A dolgozat a BCE Központi Könyvtárának nyilvántartásában semmilyen formában (bibliográfiai leírás vagy teljes szöveges változat) nem szerepel.

Budapest, 2020. 11. 18.

.....
a szerző aláírása

Contents

1	Introduction	1
2	Literature review	2
2.1	Pioneer studies from various fields of studies	2
2.1.1	Economics-related research direction	2
2.1.2	Health-related research direction	3
2.1.3	Tourism-related research direction	4
2.1.4	Migration-related research direction	4
2.2	Stock market prediction direction	5
2.2.1	The most influential researches of this area	5
2.2.2	Volatility versus return prediction	6
2.2.3	Assets in the scope of the prediction models	7
3	Research question	8
4	Biases and limitations	9
4.1	Limitations of the data	10
4.1.1	The relative data scaling method of Google Trends	10
4.1.2	Periodicity and the time frame of the observations	10
4.1.3	Daily data extraction limit	12
4.1.4	Sampling extraction method	13
4.2	Common statistical selection biases	14
4.2.1	Lack of persistence	14
4.2.2	Look-ahead bias	15
4.2.3	Survivorship bias	16
4.3	Choice of keywords	17
4.3.1	Bottom-up approach	18
4.3.2	Top-down approach	18
4.4	Capturing attention or sentiment	19
4.4.1	The definition of attention and sentiment	20
4.4.2	The practical difference between the two terms	21
4.4.3	Attention and sentiment through volatility versus return prediction	22
4.4.4	Attention and sentiment through the choice of keywords	22
5	Research design	23
5.1	Data overview	23
5.1.1	The main characteristics of the Google Trends data	23
5.1.2	The target variable: the Dow Jones Industrial Average index	24
5.2	Standardization method	25
5.2.1	The dangers of working with unstandardized data	26
5.2.2	The two-step weighing method	29
5.2.3	The rolling time frame method	30
5.2.4	The overlapping time frame method	32

5.2.5	The chosen standardization method	33
5.3	Choice of keywords	34
5.3.1	Arguments for choosing the top-down approach	34
5.3.2	The finance-related keywords	35
5.3.3	The control group	35
5.4	Development of the trading strategy	36
5.4.1	The “literature-based” approach: recreating the method of Preis et al.	37
5.4.2	The “dictionary-based” approach: introducing dictionary-based sentiment	38
5.4.3	The “data-based” approach: quantifying the relationship	38
5.4.4	The “data-based downside” approach: trading only on negative sentiment	41
5.4.5	The “data-based downside index” model: compiling the individual signals into one	42
5.4.6	Further tests regarding the strength of the collective signal	44
6	Model results	45
6.1	Technical specifications	45
6.2	Results of the model development process	46
6.2.1	Results of the “literature-based” approach	46
6.2.2	Comparison of the results of the “literature-based” and “dictionary-based” approaches	48
6.2.3	Analysis of sentiment via the “data-based” approach	50
6.2.4	Results of the “data-based” approach	52
6.2.5	Results of the “data-based downside” approach	54
6.2.6	Analysis of the economic significance of the keywords via the “data-based downside index” approach	54
6.2.7	Analysis of the collective signal strength via the “data-based downside index” approach	56
6.3	Evaluation of the performance of the final model	57
7	Robustness testing	59
7.1	Proposing the robustness test	60
7.2	The results of the robustness tests	61
8	Potential model development areas	63
8.1	Accounting for transaction costs and the bid-ask spread	63
8.2	Grammar- and trend-focused analysis of the keywords	64
8.3	Testing and documenting the implications of the standard deviation	65
8.4	Widening the scope of the assets	66
9	Conclusions	66

List of Figures

1	First problem of working with unstandardized data	27
2	Second problem of working with unstandardized data	28
3	Two-step weighing standardization method	29
4	Rolling time frame method	31
5	Overlapping time frame method	33
6	Results of the literature-based approach	47
7	Comparison of the results of the literature-based and dictionary-based approaches	49
8	Analysis of sentiment via the data-based approach	51
9	Results of the data-based approach	53
10	Results of the data-based downside approach	54
11	Analysis of the economic significance of the keywords	55
12	Analysis of the collective signal strength - first approach	56
13	Analysis of the collective signal strength - second approach	57
14	Short positions predicted by the final model	59
15	Results of the robustness tests	62
16	Analysis of the collective signal strength - transactions for the first and second approach	74

List of Tables

1	Time frame, periodicity and historical availability of the observations . . .	11
2	Keywords used in the model	36
3	U.S. sectors and their representative funds	61
4	Descriptive statistics of the search volume of the keywords	74
5	Descriptive statistics of the log returns of the DJI and the sectoral funds .	75

1 Introduction

The importance of sentiment analysis and investor attention measuring to predict stock market movements was already recognized in the late 20th century. The first milestones of the noise trading theory were laid down by Fischer Black (1986) and later Gregory W. Brown (1999). However, the vast majority of today's stock market forecasting models are still simply based on traditional methods and data sources (working with macroeconomic performance metrics, fundamental analyses, technical indicators, etc.). Meanwhile, we are living in the golden age of the social media, billions of people have access to the internet on a daily basis, unintentionally creating gigantic datasets (millions of tweets, LinkedIn posts, Wikipedia articles, etc.) characterizing the public opinion. By analyzing these mostly unstructured datasets, we will be able to quantify the investment attention and sentiment for modeling and forecasting purposes, thus addressing the human component of stock market analysis more.

Many types of research from the past decades prove the usefulness and applicability of these datasets but the question still naturally arises: which data source should we use in order to achieve the highest predictive power most efficiently. Comparing the size of the literature, the biggest support undoubtedly seems to go to the Google search volume. Additionally, Choi and Varian (2009) argue on several fronts in favor of Google. On one hand, Google data is relatively easily available for a longer time, and it is published in a quantitative form since the beginning of 2004. Moreover, they show a vast amount of applications of the data and prove the potential of it to describe economic activities, tourism tendencies, consumer behavior or even influenza-like diseases. Da et al. (2011) claim that their promising results (regarding stock market prediction) provide evidence that search volume is a quite effective predictor because it is able to capture the attention of retail investors. Later they confirmed that Google data can be also used to measure investor sentiment, especially the negative part (Da et al., 2015). A recent study of Audrino et al. (2019) directly compares the most popular forms of internet data sources including social media sites, online news portals, and search engines, from which the search volume of Google came out as one of the most relevant factors. Many other studies, that are using Google search volume for predicting stock market prices or volatility, have been getting lots of attention due to their outstanding results. One of the most acknowledged ones is the paper written by Preis et al. (2013). They are more than tripling their investment over less than 8 years with a model based on solely search volume datasets. Since that, this study is used as a framework template and a benchmark for later researches.

But is Google search volume truly that effective? Can these models also be implemented for actual real-time trading? My research intends to answer these questions by critically analyzing and following through the cornerstones and the modeling framework of the most acknowledged papers of this area. After having all the relevant data-linked and statistical

biases mapped, with a revised methodology I will try to justify the daily yields of the Dow Jones Industrial Average index (DJI) using the Google search volume of several different keywords over the past 16 years. In this manner, my ultimate goal is to test, prove and document the predictive power of search volume based forecasting, and create a model that could be utilized in actual day-to-day trading.

2 Literature review

2.1 Pioneer studies from various fields of studies

The first published papers, that stated that internet search volume is useful for forecasting certain trends, are from 2005. Ettredge et al. (2005) forecast the number of unemployed workers in the United States via Google search volume. Cooper et al. (2005) tested three potential correlates (estimated cancer incidence, estimated cancer mortality, and the volume of cancer new coverage) against Yahoo-s cancer search activity. Although the fields of studies are different, the objectives of both papers were the same: they wanted to test and calibrate the usefulness and the descriptive power of search volume against their unemployment and cancer-related datasets respectively. In the end, both of them published very promising results with high correlations and trends identified. Both of the studies stated the same: search volume is indeed a good indicator of public opinion, and since it is much earlier available than official reports, it can be used excellently in many types of surveillance systems to gain information about the present or the near-future faster. Following the success of these forerunners, in the next years, many other successful types of research emerged from very different backgrounds which all used search volume as the basis of their nowcasting and forecasting models.

2.1.1 Economics-related research direction

In economics, search-based models got popular at the time of the global recession of 2008-09 when such unique shocks appeared on the markets that were unforeseen by traditional forecasting methods. The first milestone in this area was laid down by Choi and Varian in 2009. In their paper, they forecast many types of near-term economic indicators with the help of search engine data. Their examples cover many different areas from the fields of economics: automobile sales, unemployment claims, level of consumer confidence, and even travel destination planning. Their biggest argument was that the reports of government agencies, regarding these crucial indicators, always have a reporting lag of at least several weeks (if not months), however, many private sector companies such as Google, MasterCard, or FedEx have a source of data published which could be used to assess the real-time economic activity. In the end, they used weekly Google Trends data for their research purpose. They experimented with simple seasonal AR models that included

relevant Google search data and benchmarked their results to models that exclude these predictors. They found that the former models are outperforming the benchmarks by 5% to 20%. Later in 2012 they also published an updated and streamlined version of their previous working paper which became a seminal study in this research area.

At the same time, many other successful papers emerged: Askitas and Zimmerman (2009) suggested and tested the potential of search volume in forecasting unemployment rates, Guzman (2011) utilized Google data in predicting inflation, and Baker and Fradkin (2011) used similar methods to forecast unemployment payments. Also, many papers focus on the area of retail sales and other consumer metrics. Wu and Brynjolfsson (2015) proved that microdata collected using Google Trends is also a powerful factor to forecast housing indices (that are usually only published quarterly) which can be ultimately a great indicator for near-term financial and economic collapses, as per the example of the 2008-09 crisis. Vosen and Schmidt (2011) stated and tested how search volume based forecasting performs against survey-based indicators to predict private consumption (consumer spending). Although their observed time frame was rather short (2 years), they concluded that the search volume based index outperforms its survey-based pair. They even argue that the freely available Google Trends data could replace the survey-based data collection methods in these fields due to time and cost efficiencies.

2.1.2 Health-related research direction

After the first cancer-related success (by Cooper et al. in 2015), Polgreen et al. (2008) and Gingsberg et al. (2009) proved that search data could also be utilized in the field of epidemiology to predict influenza and other similar diseases. Based on the success of these two, many other studies followed in the topic of health in the upcoming years, experimenting with even salmonella bacteria (Brownstein et al., 2009) and breast cancer (Zhang and Chen, 2010). However, the most popular field remained the forecasting of viruses and disease outbreaks, mainly due to the great intensity (or in other words attention) measuring capabilities of the search-based models. Despite the above-mentioned successes, these types of models were not yet implemented into real-time surveillance systems. According to the systematic review of health-related literature by Nuti et al. (2014), the combination of two issues is the main reason for this. On one hand, the studies are mostly geographically concentrated, based on historical health-related events of a given region or country. On the other hand, studies usually are more focused on the results and have rather poor technical documentation methodology-wise. These two reasons preclude reproducing the findings. Without such detailed documentation and the desired level of transparency, determining the actual reliability and consistency of the results is highly difficult.

2.1.3 Tourism-related research direction

Search volume based prediction is also getting more and more popular for tourism purposes. Bangwayo-Skeete and Skeete (2015) improved forecast accuracy for tourism directed to Caribbean destinations by using AR-MIDAS models containing Google Trends data against simple S-ARIMA and AR approaches. They argue that these new types of forecasts (based on flight and hotel searches) are required since demand predicting methods based on past tourist arrivals have lower and lower predictive power as the world is getting more and more globalized. According to their reasoning, this is due to people having more money, and therefore more opportunity to travel wherever and whenever they prefer. They also note that this method can be easily adapted by policymakers and business practitioners, and can be especially advantageous for planning purposes.

Yang et al. (2014) did similar time series testing with Baidu¹ data on the tourism of the Hainan province of China. Although their results were promising, they concluded many valid limitations on how effectively can be search volume used for such purposes. First of all, geography-related aspects have a much bigger impact on tourism-related searches. The majority of tourists come from foreign countries, this way one is forced to use global search volume for computational and technical purposes. However, the most popular search engines highly vary among countries, so by only focusing on one of them (e.g. Google), one can be eventually excluding millions of relevant searches (e.g. most of the Chinese population uses Baidu and not Google). Secondly, choosing adequate lag time is also crucial but very difficult. One has to be able to come up with a method to deal with different booking time spans resulting in different lags. On one end of the spectrum, there are the families, who are booking their holiday even sometimes one year in advance, on the other end, there are the young backpackers, who are sometimes basically booking a holiday on impulse for just days or weeks ahead.

2.1.4 Migration-related research direction

Based on the successes of tourism forecasting and also the recent global events, migration got into the scope of several studies. Even though some of the limitations, that was brought up by tourism-related forecasting, still stand (mainly in connection with the issue of the country of origin), the work of Böhme et al. (2019) showed some serious developments in this area as well. They recognized that the lack of migration data limits policymakers to react to the migration trends in time. Therefore, they used geo-referenced online search data of multiple flagged countries to predict migration flows, strongly outperforming any of the currently used survey-based benchmark models. This and other similar breakthroughs launched migration-focused researches (branching from mainly the tourism-predicting

1. Baidu, Inc. is a Chinese multinational technology company specializing in Internet-related services and artificial intelligence (AI). Baidu search engine is currently the fourth largest globally and the largest in China.

models due to sharing similar limitations) to appear in the past years.

2.2 Stock market prediction direction

There is no doubt that the most relevant direction in regards to our research is the area of stock market prediction. Over the past decade, more and more researches appeared in this direction as well. In this review, I tried to categorize all the most relevant studies, in order to gain a deeper understanding of how past articles utilized Google Trends data in their frameworks, how they tried to approach a similar problem from different angles.

2.2.1 The most influential researches of this area

First of all, I am going to present three of the most influential studies of this area, which are among the most cited and most acknowledged studies of stock market prediction via Google search volume. Although their approach is rather different from each other, all three papers brought relevant findings and made a huge impact in this field of study.

One of the most cited papers in this area was written by Tobias Preis, Helen Susannah Moat, and H. Eugene Stanley in 2013. Huge amounts of data are now being generated through the extensive interactions of people through the internet, automatically documenting collective human behavior in a new fashion (done by private companies such as Google). Their hypothesis was that interactions through the internet not only reflect the current state but provide insights into future trends of the economy as well, which are usually first presenting themselves through stock market movements. Therefore if we have a structured database of these interactions (like Google search volume), we would be able to forecast stock market movements for a short time ahead. To prove their theory they observed weekly time periods from 2004 to 2011 to compare the price changes of the Dow Jones Industrial Average index (DJI) and Google Trends data for certain keywords (characterized by the closing price of the first trading day of the week and the aggregated Google queries respectively). They tested their hypothesis via a simple buy and sell trading strategy based on the search volume of 98 keywords. During this process, they found that they could achieve the highest yield with the keyword “debt” resulting in a 326% overall cumulative yield in less than 8 years. Their study is highly popular and acknowledged in this area of research, many later articles cite their findings and use their framework and methodology for financial forecasts, and even for predictions in other fields of studies.

We also have to definitely highlight the studies of Zhi Da, Joseph Engelberg, and Pengjie Gao from 2011 and 2015. The first paper of Da et al. (2011) is focusing on measuring investor attention using Google search volume. They run tests on the Russell 3000 stocks from 2004 to 2008. They compared the search volume of keywords of company names and tickers to other attention-measuring indicators. They have found that although the

search data is correlated with these existing proxies (like turnover rate, extreme returns, information gathered from financial news, etc.), it captures attention in a different way. They also provide evidence that the search volume can be regarded as a direct, objective, and relatively easily quantifiable measure of the attention of the retail investors, which has been highly difficult to capture before. They also experimented with event studies on IPOs.

It is quite surprising that the second paper of Da et al. (2015) approaches the problem from a very different angle. Here they have used search volume as a proxy for a market-level sentiment of households, thus for the overall sentiment of retail investors. They chose keywords that express typical household concerns towards the economy (such as “recession”, “gold”, “bankruptcy”, etc.), and by aggregating their search volume they have created the “Financial and Economic Attitudes Revealed by Search” (FEARS) index. With their data-driven approach, they have not only predicted market returns but anomalies related to arbitrage trading, volatility, and fund flow as well. Among others, they have also found that negative sentiment, thus the search volume of keywords that the regression-based model signaled as having a strong negative relationship with the market, is more reliable for forecasting the stock market.

The findings of these papers, especially the monumental yield of the “debt” keyword of Preis et al. (2013), attracted a lot of attention to this new aspect of stock market prediction. Due to the fact that this research area was relatively intact beforehand, many different adaptations and research directions began to appear afterward, mainly based on the above-described models. During my research, I reviewed several of the later papers as well, in order to gain a deep understanding of all the potentials and limitations of this stock market modeling approach. Most of the findings of the reviewed works will be introduced at later stages (during the overview of the modeling biases and the methodology), now I will summarize and categorize the main concepts of how researches approach this topic by introducing a few relevant examples.

2.2.2 Volatility versus return prediction

The most popular direction is to forecast stock market volatility via search volume. A great example of this type of researches is the paper by Dimpfl and Jank (2015). They incorporated search volume into many time series prediction models to test on realized volatility indicators of the U.S. market. They found that augmenting the base models with the search queries leads to more precise results, especially concerning the long run and the time spans with abnormal volatility. Thus the utilization of search queries could be used even in real-time to forecast future volatility. Later works, such as the paper of Xiong et al. (2016), apply more developed models using neural networks, which are even outperforming the linear and AR models including search volume. These successes also

show the potential of deep learning methods trained with search queries in the presence of strong noise.

The other direction aims to predict stock market returns with the help of search volume. For this direction my earlier observation is particularly true: most of the researches built their models entirely from the model of the seminal paper and they use their results as benchmarks. Kristoufek (2013) proposes a novel approach to portfolio diversification using Google search volume to maximize return. The diversification is based on measuring the popularity of a stock through search queries. He penalized the popular stocks with lower portfolio weights, while he brought forward the peripheral stocks with bigger weights. By adding this element to the trading strategy of the paper of Preis et al. (2013), he could beat the out-of-sample returns of the benchmark DJI index by 38%. Later studies also reference these works. Zhong and Raghieb (2019) further developed the strategy of Kristoufek by building in an adaptive keyword selection mechanism, resulting in an almost 500% cumulative return in 10 years.

We must note that the vast majority of the researches, as I already mentioned, deals with forecasting volatility. The idea of that is consistent with the noise-trading theory of Black (1986): search volume translates to the individuals' interest in the aggregate stock market, and this is reflected in excess volatility. Forecasting returns seems not that straight-forward (and for this reason not that popular either). Even if a correlation can be assumed between the search volume of a keyword and a financial asset, the changing search volume of a given keyword does not necessarily indicate the direction in which returns might change.

2.2.3 Assets in the scope of the prediction models

Researches capture investment attention and sentiment shaping the economy usually by forecasting comprehensive financial indices like the DJI, S&P500, or the Russel3000. They can do this directly through search data of carefully chosen keywords, as we could see from the listed literature so far. There are also initiatives to forecast the price changes of indices by forecasting its stock components one by one. The main argument of Da et al. (2011) is that excess returns, which could be captured by the search volume prediction models, are highly event-based: the events of individual stocks (IPO, bankruptcy, etc.) spill over and cause the biggest drops and increases in the aggregate index price. Therefore in their research, they are experimenting with event-study methods to determine the correlation between search volumes of tickers of given stocks for their IPO period. They conclude that search volume contributes to the large first-day return and the long-run underperformance for their sample of IPO stocks.

Google search volume is also a popular forecasting method in other markets as well. Li et al. (2015) utilized it to forecast and measure the price volatility of crude oil and managed to improve the results compared to competing time series models. There were even attempts

to forecast the Bitcoin price. However, as per the findings of Urquhart (2018), in the case of Bitcoin the causality is rather the other way around: daily realized volatility and volume significantly influence the next day's attention (search volume) but vice versa the predictive power is insignificant.

Last but not least, many of the researches, either predicting volatility or return, only use internet search volume as an additional input to their models. These models require heavy computing power since they are collecting sentiment and attention data from many different sources. Audrino et al. (2019) developed such an attention index by combining signals related to individual stocks and stock market indices, using text data from two social media platforms (Twitter and StockTwits), financial news articles (obtained from RavenPack News), and volumes of two search engines (Google and Wikipedia). After developing and testing the index, the results of both the in-sample and out-of-sample tests showed that basically, two components are the most relevant predictors from the collected datasets, one being the Google Trends search volume (the other one is the StockTwits news articles). Thus further proving the relevance of the Google search volume for such predictions.

3 Research question

The paper of Preis et al. (2013) is considered as a popular methodological template by many of the later researches (working with Google search data) due to its marvelous results. It is not followed in the field of finance, but in other fields of studies as well, such as by researchers forecasting macroeconomic indicators, tourism, spread rate of diseases, etc. The same could be highlighted regarding the works of Da et al. (2011 and 2015). Unfortunately, they are readily cited as a ground thesis by others without looking deeper and changing anything in the key elements of the modeling frameworks.

The work and the achieved results of these authors are certainly ground-breaking and inspirational, however, we should not take the methodology as given, even if it promises high results (like the outstanding yield of 326% of the flagship keyword 'debt' of Preis et al. (2013) in under 8 years, or the yields of a slightly modified version of Zhong and Raghbi (2019) scoring almost 500% in 10 years). These exceptionally huge yields should raise first serious questions in connection with the reliability and credibility of the modeling frameworks that produced them. In my opinion, indirectly this can even hinder further methodological developments in this area since many other later papers might not get the same spotlight if they can't achieve something of similar magnitude. To combat these issues, the goal of my research is twofold but closely connected.

First, I aim to thoroughly analyze the modeling framework of the papers introduced in the literature review. The main question is whether the outstanding results described there

are reliable and well-grounded, or the huge yields and the striking conclusions seem more of a result of overlooked biases and untouched limitations. During my careful inspection, I will be focusing on two main factors. On one hand, I would like to map the key areas of the models of the most influential papers, where limitations or biases could have been overlooked, and which might significantly alter their final results. On the other hand, I intend to find possible rooms for improvements, which can ensure that such a model is developed that could be utilized in actual real-time trading at a later stage. In order to have a comprehensive and extensive picture here, I will be also considering other relevant studies with best practice examples and important lessons to be learned. The methodology of my model will be constructed based on the observations made there, which leads to my second hypothesis.

In a wider sense, my research also seeks to answer the question, whether Google search volume truly has such a huge predictive power to forecast stock market movements. To provide proof of the utility and practicability of such models, I intend to present results based on a well-documented, structured and prudent methodology. By doing so, I sincerely hope that it inspires others to similarly challenge the status quo of this field of study, thus eventually contributing to the further development of the highly promising search volume based modeling approach.

4 Biases and limitations

Prediction about the future is prone to many kinds of methodological biases. Therefore setting up a correct and prudent methodological framework is essential. If we (accidentally) overlook some biases, they could considerably affect the model's reliability and question the significance of its results. This can be especially dangerous because these types of missteps often positively alter the results, but in reality, our conclusions derived from the model might be meaningless. These methodological weaknesses can be avoided if we thoroughly examine the underlying datasets, and come up with ways to handle all its limitations. Another common bias arises from how effectively we are able to evade exploiting any information which is available now but was not in the past. There are numerous other specific biases that we have to recognize early on in order to have the correct framework for a reliable model.

Therefore in this chapter, I will be closely inspecting modeling cornerstones (one by one) to see how the authors of the former papers dealt with them. I will be also introducing alternative solutions, backed up by the findings and opinions of other famous and popular researchers in this area. This way I will be able to come up and lay the groundwork for my methodology, which complies with the aspects detailed here on a theoretical level. Among these biases, I will introduce some enhancements regarding the limitations of the

datasets, common statistical selection problems, the process of choosing keywords, and terminological flaws.

4.1 Limitations of the data

Meanwhile, Google aims to supply a clean and structured search volume dataset it has to apply several limitations for different reasons, such as preserving user anonymity, data security, etc. Therefore the first and foremost point we have to concern is the main characteristics and the related limitations of the data imposed by the provider.

4.1.1 The relative data scaling method of Google Trends

Google Trends provides an aggregated (therefore anonymous) analysis of the volume of given Google search queries across various regions and languages over set periods of time. It provides not the absolute but the relative search volume of a given keyword in a given time frame. These relative values are scaled between 0 and 100: the local maximum of the search volume within the time frame gets 100, and the rest of the values are scaled relative to this local maximum value. The relatively scaled search data is only applicable and pertinent to the given time frame and region that it was originally requested for. Therefore, if you request data for another time frame or another location, the relative scaling process is repeated according to the data points of the query with the modified input variables. On the grounds of this fact, the result of two data requests, with different time frames and/or locations as input variables, are not comparable, at least not without applying some kind of standardization procedures (for more details please refer to Section 5.2).

Google makes an enormous effort to provide as clean search data as they possibly can. They apply a very robust and complex algorithm, multiple other mechanisms, and cybersecurity measures to monitor and filter out irregular and abnormal search activity, search spam, and other kinds of suspicious user behavior. Adhering to database management principles, they don't retain search data for past periods, where only a negligible amount of searches happened from only a few users. These periods are later assigned a 0 automatically, after wiping out these records from the database. The algorithm is also indifferent towards obvious spelling errors, special letters, characters, hyphenation, and punctuation marks, thus terms containing these are still contributing to the search volume of the same keyword.

4.1.2 Periodicity and the time frame of the observations

Search volume data is available on Google Trends since the beginning of 2004. You can request data on different time frames, and the length of this time frame determines the periodicity of the data. Table 1 summarizes the connection between the different time spans, their periodicities, and their historical availabilities.

Time frame	Periodicity	Historically available
past 4 hours/1 hour	1 minute	No
past 1 day	8 minutes	No
past 7 days	hourly (1 hour)	No
1 day - 90 days	daily (1 day)	Yes
90 days - 5 years	weekly (1 week)	Yes
5 years+	monthly (1 month)	Yes

Table 1: Time frame, periodicity and historical availability of the observations

Source: Own table

As we can see from the table above, one can get even really high-frequency data, but Google only provides historical search volume datasets for certain time spans. It goes without saying that in our case we would need a huge amount of historical data in order to conduct comprehensive testing of possible trading strategies over years (if not decades). Therefore our only options are datasets either with daily, weekly, or monthly periodicity. Although Tetlock (2007) was using newspaper articles as input for constructing his sentiment factor (and not Google data), he provides another reason why one could avoid dealing with intraday high-frequency data. He ran comprehensive robustness tests and a sensitivity analysis whether the predictive power of his pessimism factor is concentrated in the opening hour returns, or in the after-hour returns, or if it is dispersed uniformly throughout the entire trading day. His results prove that the changes in the market returns, following a pessimistic article, are not concentrated after the release of the specific information, but can be observed throughout the entire trading day. He even adds that remnants of this effect can be found in the following days as well, further confirming that due to the reaction time of the investors there is no need to touch intraday data.

One should also consider other technical limitations of the Google Trends engine because we can't directly set the periodicity, however, we can only set the time span which determines the periodicity of our dataset (as shown in Table 1). If we want to request data for longer time intervals, the periodicity of data becomes longer as well: within 90 days we get daily data points, from 90 days to 5 years we can get weekly data points, and above 5 years we are provided with monthly data points. Additionally, there is a small delay for publishing the data, for Google to have enough time to collect, clean, and structure the search volumes. Daily data can be collected 36 hours prior to the search, weekly data (always ranging from Sundays to Saturdays) is available on Sundays, and monthly data is published always 2 days after each month-end (Google Trends, 2020).

Choi and Varian (2009) are some of the most famous forerunners, who first performed 'nowcasting' with monthly data downloaded via Google Trends. They exploited that in many different industries and fields (motor vehicle sales, unemployment statistics, travel statistics, consumer confidence index) monthly data is only available with at least a

two-week-long lag, meanwhile Google data is available after only two days. Therefore, by using the Google search volume of certain related keywords, they are able to ‘forecast the present’ (nowcast) much sooner. They have found that simple AR models, that have Google data in themselves, outperform significantly AR models excluding it. Unfortunately, while monthly data proved to be appropriate for forecasting/nowcasting in these areas, it is not preferred in connection with capital markets, where many frequent data points might be required. In the paper by Preis et al. (2013), we can see that the authors used weekly periods for this purpose, and many of the later researches in this area followed this example. By using a weekly periodicity over a monthly one, data points are more frequent, however, we would still be predicting aggregations of daily price changes. If we want to use Google data to its full potential, we should obtain daily search volume data. With daily search volume, we would be able to directly forecast daily price changes (daily log yields) on the capital market. However, there are two issues with this approach we should first consider.

First, the most ideal case would be if we could apply a one-day lag, meaning, that with the search volume of day t , we can forecast the price change between day t and $t + 1$, thus the daily log yield observed on day $t + 1$. This is, in theory, a viable option with historical data, however, in practice (in actual trading) this could never be utilized because Google is only publishing its data 36 hours after the end of the given day. Therefore to avoid any biases, a minimum of three days lag should be applied (the original one-day lag plus the 36 hours of waiting time, rounded up) if we stick to the usual calculation method of daily returns (meaning, that relative portfolio changes are derived from the daily adjusted closing prices). Secondly, we also have to pay attention to the fact that daily data is only available in 90 days spans, and independently downloaded daily data is only comparable within that given time span. Therefore, if we want to compare data across different windows, we should definitely apply a standardization technique to be able to scale the data onto the same spectrum.

4.1.3 Daily data extraction limit

The difficulty of data collection and structuring does not lie in the querying part but in overcoming the limitations imposed by Google. We call one query a data request for a given keyword with a preset location and time span. Repeated queries (data requests spammed right after each other) from a single IP address are disregarded by Google to suppress potential threats arising from automated data-scraping bots and other algorithms. Therefore, Google set 200 as the daily limit of repeated data extraction requests. Google Trends does not recognize several data requests as spam (and does not apply the 200 limits) if at least one minute break takes place between each of the requests. Moreover, for the same keyword and time frame, you will get the same results if you are querying

within the same hour. All of these limitations make the querying of the realizations of the datasets for the same time frame very cumbersome (the reason, why more realizations are needed, is mentioned in the next section). While respecting the legal rights of Google, these issues are only solvable by using more capacity (more computers and networks) for downloading, which is for individual researchers not readily available. The other option is to manage and plan data downloads strategically.

4.1.4 Sampling extraction method

I have also found during querying that search volume data changes slightly over time. This is due to Google’s sampling extraction method, meaning that Google does not provide the relative volumes based on its overall data, but first samples it, and gives the results of the query based on the sample. Although most of the researches neglect (or don’t even realize) this issue, we should definitely investigate it, since it might render our data totally useless: if the sample of searches is not representative, it would mean that the different samples are significantly different from each other, and performing any kind of forecasting with it would be a waste of time and effort. Fortunately, the documentation of Google Trends (2020) confirms the representativeness of the sampling method, therefore, the different realizations of downloaded datasets for the same time period should not be significantly different from each other. They also state that they are using sampling due to data protection and/or capacity-sparing purposes. This way they are able to provide representative data within seconds for multiple users simultaneously.

The paper of Preis et al. (2013) uses the average over three realizations of its search volume time series, based on three independent data requests. Averaging could be a valid solution, but it forces us into a paradox situation because it fails to address the real question. Downloading and processing only a few realizations is although very capacity-friendly (and also not that time-consuming), it does not help if the different realizations are far too different. Parallel, if the different samples are almost identical, averaging over them would be a redundant extra step. Therefore, instead of applying “quick fixes” to our dataset, we should be focusing on the core problem: confirming how identical different datasets are to each other.

D’Amuri and Marcucci (2017) were examining and forecasting the monthly U.S. unemployment rate based on the search volume of “jobs” and many related keywords. They have found that cross-correlations between independent realizations of these keywords for the observed 10 years are always above 0.99. For the calculation, they analyzed 24 realizations of the monthly dataset of one keyword. Böhme et al. (2020) used monthly Google searches to predict migration waves towards Europe. They used 67 different migration-related keywords (each in three languages: English, French, and Spanish). They have arrived at a similar conclusion regarding the cross-correlations of keywords (although unfortunately,

they did not attach a detailed methodology to support it).

As I have already mentioned, the sampling issue is rarely considered in the literature. The conclusions of the above two papers are enticing, but we should note that both of them were performed on monthly datasets. To gain direct evidence, similar testing would be required on daily datasets, precisely tested for our keywords, since correlation might vary based on the size of the datasets for different keywords. However, this would demand huge computing capacity, since downloading one daily realization of one keyword from the beginning of 2004 would take more than 4 days due to the extraction limitations. Additionally, we should not forget that Google Trends creates monthly data points by aggregating daily data points. All in all, we should not assume directly the the 0.99 cross-correlation applies not only to the monthly but to the daily datasets as well, because presumably, the cross-correlation of different realizations of the same daily dataset is somewhat lower. However, based on the findings of the above two studies, and also considering the statement about sample representativeness from the documentation, we can still move past the sampling issue.

4.2 Common statistical selection biases

4.2.1 Lack of persistence

Researchers generally agree that forecasting with time series should use out-of-sample testing rather than an in-sample one, in order to maintain persistence in the modeling framework. The argument has two major points. First, in-sample errors are likely to understate forecasting errors, so we are basically overfitting our data. Since the future might hold nuances that might have been not revealed in the past, the results of in-sample testing due to overfitting can be very misleading. Secondly, the best in-sample fit does not guarantee the best post-sample fit. By dividing the data into training and testing periods (and by using the training period for fitting the model and the testing for assessing the performance), we can simulate real-time assessment without having to bother with its practical limitations (Tashman, 2000). According to the thoughts of Leinweber (1995), we have to be very careful what we ask for when training prediction models because we are very likely to get it, meaning, we unintentionally tend to overfit models to gain the desired results. In addition, testing the process to see if we can reproduce models of similar quality, using data we think is random or on a time frame that was not part of the training-phase, can be a sobering experience. Performing only in-sample training and testing is one of the most severe examples from the family of confirmation biases.

The aim of Preis et al. (2013) was probably not to create a functioning training strategy but to assess the relationship between Google search volume of certain keywords and financial returns. However, it is very surprising that no out-of-sample test was performed to support the results of the study, and thus the persistence of the model. Their model

was fitted on the same data that they used to measure the performance of it. By reading the article for the first time, the huge overall yield of the ‘debt’ keyword throughout only 8 years is very eye-catching. However, after taking a deep dive, and realizing that the published results stem from an in-sample test, it is advised to be skeptical, and question the strategy’s persistence.

As to my knowledge, many other papers in this field of study, even if they conduct an out-of-sample test, usually only validate their model by dividing their time series into one training and one testing period. Even though this technique is more appropriate for cross-sectional analysis. Dividing your dataset like this gives only a one-time proof that the model works, but it fails to prove overall persistence at all because it assumes that model parameters are constant even over longer time periods. When analyzing time series data, especially one that considers the economic environment, it might not be reasonable to assume this. Therefore, it is encouraged to use a rolling or expanding validation method for measuring the performance. If it turns out, that parameters are indeed constant over time, then the results of the rolling or expanding windows should not be different. On the contrary, if parameters change at some point, then these techniques are able to capture that instability, and we can train the model accordingly (Zivot and Wang, 2003).

The main advantage of the rolling training method is that it allows the model to react to sudden structural changes fairly quickly. On the other hand, our model is not able to learn from events outside the rolling window, which can deprive a huge pool of valuable historical data from the training cycle. In the case of economy-related forecasting, this can be above all hindering, because we could miss out on rare but critical past events, such as financial crises, market crashes, IPOs, or mergers and acquisitions. However, where the rolling training method lacks, the expanding training method excels, because it considers the whole available historical dataset when training. Unfortunately, this might also affect the reaction time of a prediction in a negative way compared to the rolling one. But eventually either a rolling or expanding method should prove more reliable and robust than validation methods with pre-fixed training and testing time frames (not to mention in-sample tests).

4.2.2 Look-ahead bias

The look-ahead bias is one of the most common selection biases that distort statistical analysis, usually rooted in the incorrect selection of samples. This bias occurs when one uses information that would not have been available during the period used in the analysis. Despite avoiding using future data sounds obvious, it is a typical error, mainly because researchers add this bias unintentionally to their frameworks. There are a few ways and checkpoints with which one can avoid this particular bias.

As a first step, a researcher should get to know the original database and the source of it

in the best possible way. He or she should be able to answer questions like when and how frequently the data is usually published, are there any late adjustments, and what other important characteristics does the data have. Most importantly, one should confirm before touching the dataset that no biases, especially look-ahead bias, are involved inherently. Usually, this step is neglected, such characteristics of publicly available data are rarely checked and justified. After this first step is successfully concluded, and one begins the model building process, there are a few widely accepted cautionary measures to be taken. Starting with a small sample, and building a model using rolling or expanding training and validation methods usually takes care of the look-ahead bias. There are also randomized cross-validation methods for time series prediction especially designed to find any trace of such biases. Leaving out a short testing period from our dataset for self-validating can also work as a solution (Walimbe, 2017).

We can see from the above examples that there are numerous statistical resolutions to escape look-ahead bias when modeling, however, handling the data-specific (the source-specific) look-ahead bias is always a very specific task and is totally up to the researcher. In the case of Google Trends data, we already mapped all of its most important characteristics and limitations in Section 4.1. The fact that the data is scaled relative to the local maximum point onto a scale of 0 to 100 seems to be a serious source of look-ahead bias. And this is even before trying to cover longer time periods by binding together independently downloaded datasets. Linking together these realizations without a careful and statistically correct standardization procedure can be another possible source of look-ahead bias. A deeper and highly data-specific analysis of these two issues and a development process of possible solutions will be presented in Section 5.2, the purpose of this chapter was to highlight the fundamental need for such measures.

4.2.3 Survivorship bias

In a broader sense, survivorship bias is a selection process error caused mainly by the focus of visibility: only a part of the observed population is selected into the sample due to a few criteria that give them visibility, while others are overlooked because of not fulfilling it, thus not being visible enough. In finance, this is a common bias, because the comparison of performances and other indicators is fundamental, especially in the case of the stock market and company valuation-related topics. Researchers tend to always focus on investigating the extremities, in the case of researches considering stock market comparison and prediction usually the outliers of the upper quartiles prevail. Therefore, the survivorship bias not only can lead to false conclusions, but to extremely optimistic expectations as well (Brown et al., 1992).

When talking about Google search volume prediction, I believe there is a direct interpretation of this issue. In Section 3, I have already mentioned what striking results were

achieved by a few of the researchers in this field of study. However, in Section 4 (and later in Section 5), after a thorough analysis of the methodology followed by some of those papers, I concluded a great some of undocumented decisions and statistical oversights. Unfortunately, due to the survivorship bias, many people are blinded by the huge results, and they fail to take a few more steps to check and validate the methodology behind these model outcomes. Consequently, other papers, which may have completely correct modeling frameworks, are overlooked, because they could not realize such high cumulative yields with those. Survivorship bias, therefore, results in a paradox situation: success tends to be judged based on the results themselves, and not on the methodological achievements.

We can also see some traces of the survivorship bias when looking at the choice of keywords. First of all, researchers tend to choose keywords subjectively, based on the selection of previously successful studies. They presume that once a keyword was proved to be promising, it will be in the future as well. This unintentionally introduces permanency into the frameworks, because then the models are run with some parameters (the keywords) assumed to be constant over time. Whether this assumption stands, should be tested during the model development process by applying a fully adaptive, data-driven methodology. The relationship of the search volume of the keywords with the market will be covered in Section 5.4.3, while the actual selection process of the keywords is presented in the following chapter.

4.3 Choice of keywords

Choosing keywords is a crucial step when we predict with search volume data. In most of the studies the method of selecting keywords is not taken seriously enough, although by neglecting this issue, the whole research procedure could be significantly biased and mislead. In most of the studies, selecting keywords is only based on some personal considerations of the authors without any extensive reasoning. In this chapter, I will be focusing on the few exceptions, but even these studies introduced very different ways to address this issue, thus there is no best practice method developed yet.

Among the studies, which addressed the issue of choosing keywords, we can find ones that are predicting volatility and not return. Moreover, some studies are only using the Google search volume as one of the inputs among many other indicators (like Twitter data, financial news analysis, etc.). However, for the purpose of choosing keywords, the listed differences are irrelevant, the focus should be on the technical mechanisms. Although, it is important to note that all of these studies are forecasting US indices (typically either the Russel3000, S&P500, or the Dow Jones Industrial Average Index). In my opinion, the literature can be divided into two categories based on the selection processes: the bottom-up approach and the top-down approach.

4.3.1 Bottom-up approach

In this approach index forecasting is performed more from a technical aspect, researchers decompose the index into its components, the shares, and try to forecast the index by capturing and aggregating the forecasts of the individual securities. Typically in this case company-level, company-related words and phrases are used as keywords. However, there are certain limitations with this type of approach that has to be accepted (or dealt with), as per the study of Da et al. (2011), one of the most acknowledged studies of this area. According to their train of thought, people are using either the company names or the stock tickers to search for individual companies.

Identifying search frequencies by names might be rather problematic for the following two reasons. First, one might search for these companies unrelated to any trading activities (such as searching for Microsoft because of Excel-related questions). The bias here is more severe if the given company name has several meanings (like Apple or Amazon). Secondly, there is no clear unique identifier for a given name: in the case of some firms, one can choose from multiple variations of its name to search for (such as you can use AMR Corp, AMR or even AA to search for American Airlines).

Searching with a ticker is certainly less ambiguous, therefore, this became the chosen method by Da et al. (2011) eventually, but they note a few things here to consider as well. First of all, via tickers mostly only trade-related searches will be captured. This does not have to be a disadvantage since these are the searches that will result in actions that have a direct effect on the market. However, due to Google's limitations, when we have this info in our hands these trades probably already happened so we can be late to react to them. Moreover, we have to be cautious because some of the tickers can have generic or ambiguous meanings (such as GPS, DNA, BABY, etc.). They call these the “noisy tickers” because usually, their search volume is abnormally high from attention unrelated to the underlying stock. In the Russel3000 index, they flagged about 7% of their stocks as noisy ones, but since this flagging was done based on their subjective consideration, they left these stocks in their sample (they confirmed that their results are robust enough if they were to exclude them as well).

4.3.2 Top-down approach

I named the second type of search word selection method as the top-down approach. In this type, researchers try to capture the movement of the market by concentrating not only the focused company- or industry-related but the whole attention of the public. They are usually using finance-related generic words which supposedly have a strong economic sentiment. This approach was also used by Preis et al. (2013). They analyzed the performance of a set of 98 search terms, which were selected based on their own consideration, or were suggested by the Google Sets service (this service was since dismantled by Google). Most

of the terms (like "debt", "crisis", etc.) were related to the concept of trading, intentionally introducing some financial bias.

Most of the other articles also tend to use keywords that were chosen based on the authors' consideration. Granell and Carlsson (2018), the authors of one of the (somewhat) counter-examples, note that finding a decent method to choose adequate search words is particularly difficult because there are no stated theories or best practices on how keywords can be chosen for prediction or analyzing casualties. In their study, they chose the 20 terms (such as "debt", "stocks", "shares", etc.) from "20 English words for finance you simply must know" (by the popular business blog FluentU). However, in the end, they also decided to add a few terms based on their own consideration as well (such as "crisis", "S&P500", etc.).

Another best practice example is the research conducted by Perlin et al. (2016). They have a rather critical opinion on how the choice of keywords was performed so far by other researches in this area, therefore, they tried to innovate the process. First, they used the internet finance dictionary Investopedia as a source, they extracted all the words found in the dictionary, resulting in almost 15,000 unique terms. As a second step, they analyzed four finance books: two popular academic ones and two Amazon best-sellers from the finance section. They calculated the number of appearances of each of their previously extracted unique terms in all of the four books. The 15 terms with the most occurrences (including "finance", "capital", "value", "debt", etc.) were chosen as their list of keywords for forecasting.

Challet and Ayed (2013) specifically criticize the choice of keywords by Preis et al. (2013). They say that it is natural to think that keywords related to finance and capital markets have a bigger chance to perform better in forecasting financial indices. There is no problem at all with adding a financial bias when choosing keywords, however, this bias needs to be controlled with a set of random keywords that are unrelated to finance. This aspect is highly neglected by other researchers too, although it could be only due to luck or model miscalibration that their chosen financial keywords are high-performers. Therefore they concluded a study simulating the methodology of Preis et al. (2013), but added 400 control keywords (including medical conditions, illnesses, types of classic cars, and titles of arcade games). With simple t-testing they concluded that although there are finance-related keywords that perform well, there are lots of control words that have similar results, thus confirming their previously stated concerns.

4.4 Capturing attention or sentiment

Most of the researches presented in my literature review uses the sentiment and attention terms as synonyms, only a few make a direct distinction between the two terms, and even fewer take the extra step to handle these concepts separately. It is also surprising,

that there are actually no clear definitions in the literature of how these terms could be translated to the field of Google search volume prediction. The lack of distinction not only appears in the terminology on a theoretical level but in the methodology on a practical level as well, because there are no generally agreed-upon methods to capture these. In this chapter, I will be focusing on making the thin line between the two terms clearer, by describing my observations following the literature review.

4.4.1 The definition of attention and sentiment

In my view, the change in investor attention is represented directly through the actual change of the search volume. The relationship is quite straightforward: a search volume decrease for a given keyword means that the attention directed towards a certain market element (company, industry, or the whole market) decreases as well, while an increase in search volume equals increased attention. The idea of sentiment is more difficult to grasp because it is more complex. Sentiment is represented by not only the change in search volume but the strength and the direction of how this change relates to the market movements. This means that measuring attention is a precondition of measuring sentiment. Sentiment, therefore, comes from the combination of attention and the relationship of this attention with the market, the latter characterized by mainly its strength and direction. These two terms, attention and sentiment, are often mixed up throughout the literature. In my opinion, therefore, each model functions like a sentiment-capturing model eventually, even if only the attention-capturing part seems to be relevant. The reason behind this statement of mine is that attention does not imply any relationship with the market in itself. This way, even if we don't measure the relationship directly (even though in my opinion we should do so), we make an assumption of what this relationship might look like. As an example, if we say that the increased attention, thus the increased search volume of a certain keyword, will result in increased market returns, we assume a strong positive relationship of the attention of this keyword with the market. In my opinion, this is equivalent to stating that the keyword has a positive economic sentiment. Please be aware, that by accepting this logic, there is an alternative way of observing positive economic sentiment: decreased attention, thus decreased search volume, resulting in decreased market returns. To sum up, we have a keyword with strong positive economic sentiment when the attention and the market return strongly positively correlate. Meanwhile, our keyword bears strong negative sentiment if the attention and the market move almost completely in the opposite direction. Additionally, if the relationship of the attention and the market movements don't correlate, we are talking about a keyword that has no (or neutral) economic sentiment.

4.4.2 The practical difference between the two terms

To make the whole above-described concept less abstract and more digestible, I will bring actual examples from popular and influential researches. Please note that these strategies will be discussed in details later in Section 5.4, when constructing my own trading strategy, here I am only using them for the purpose of showing the difference between the two terms, without going deeper into the actual technical specifications and statistical background.

Preis et al. (2013) don't specifically position their study on the spectrum between attention and sentiment beforehand. Their trading strategy utilizes a historical moving average: if the current search volume of a given keyword is higher than the moving average of the last few weeks, then they are taking up a short position in the DJI, and vice versa for a long position. If we want to pair this strategy with the concepts I described above, we can see that this strategy focuses completely on capturing investor attention, and it totally neglects to measure relationship, thus sentiment too, rather it assumes a direct determination of how the attention relates to the market. A search volume increase, thus an increase in attention, is directly linked with a strong negative relationship because a short position is taken up in this case. While (following their logic) a decrease in attention is paired with a strongly positive market relationship, thus a long position. Although they are only focusing on capturing investor attention, they are automatically assuming that the market movements relate negatively to attention, with a strength that is constant over time. So they are assuming that each of their keywords bears a strong negative sentiment towards the market. That is an entirely separate question whether this is intentional or correct to assume.

The "In search of attention" article of Da et al. (2011) is one of the first acknowledged researches that make a clear distinction between sentiment and attention. Although they avoid clearly defining the terms, they mention that attention is a precondition of having sentiment. In this study the authors are focused on capturing investor attention: they are using company names and tickers as keywords, thus the change in the search volume of these would be the proxy of the investor attention directed towards these companies respectively. They are conducting an event study, examining the time frame of IPOs, therefore they make the assumption that the retail investor attention and the market are positively related. Although it sounds rather abstract, they indirectly presume that company names and tickers as keywords have a positive sentiment when we are observing IPO events.

Later Da et al. (2015) in the article "The Sum of All FEARS Investor Sentiment and Asset Prices" present a totally different approach, which is based on capturing investor sentiment (as the title already suggests). In this extensive paper, they put a high emphasis on how sentiment, the strength and the direction of the relationship of the search volumes with the market, could be measured and built into a trading model. They don't handle attention here

explicitly, presumably because in their former study they state that capturing attention is already a prior condition of being able to measure sentiment.

4.4.3 Attention and sentiment through volatility versus return prediction

From the above-stated observations, one can discover similarities between attention and sentiment, and volatility and return prediction. It seems that attention actually indicates volatility in the market. Theoretically, volatility means the standard deviation of market returns, so just by itself, it does not inform us which direction the market is moving, it only signals a change of market prices. To link the volatility to returns we need extra information: how the volatility relates to the market returns. This is very similar to the logic of how the concepts of attention and sentiment were connected.

This means that for measuring volatility capturing investor attention should be enough, since volatility is the actual depiction of the change of investor attention, which, following our definition, can be observed through the change of search volume. When we are predicting returns, we are also starting here. The attention “first” creates volatility in the price, but we are in need of a further step to be able to assess the strength and the direction of the change in price. This further piece of information will give us investor sentiment, and so, the capability of forecasting market returns. In most cases, when the concept of investor attention and investor sentiment is mixed up in this field of study, researchers are actually falsely linking together volatility and return prediction. And as we generally don’t blend together these two fundamental financial concepts, we also don’t want to do that here. This is another important reason to have a transparent and clear distinction between the definitions and the measurement methods of investor attention and sentiment.

4.4.4 Attention and sentiment through the choice of keywords

Based on the literature so far, we can also see an indirect link between the duality of attention and sentiment capturing approaches and the keyword selection methods. The aim of the bottom-up selection method is exactly to find keywords that can capture focused investor attention. That is the reason why usually tickers, company names, or industry-related terms are chosen here as keywords. Unfortunately, the closer investigation of the sentiment is usually skipped here, researches tend to hypothesize that these terms, thus their search volume, have an already strong relationship and a clear direction towards the chosen segment of the market. While on the contrary, the top-down keyword selection method seems to be used for sentiment-orientated models more. In this case, the goal is to choose keywords that are generic and economically relevant since one should be focused not only on capturing the attention but on studying the correlation and the intensity of the relationship between the search volumes and the market movements.

It should be highlighted that I am not stating that the method of choosing keywords unambiguously determines whether the model would be focused on rather capturing attention or sentiment. The above statements are more based on a trend that can be observed through the literature of the Google Trends prediction. In my opinion, no matter the keyword selection method, both attention and its relationship with the market should be handled in a model, even if the latter one seems to be evident. Therefore, whether a model becomes more focused on investor attention or investor sentiment in the end, should not be the consequence of how and what keywords have been chosen. It should be a result of what statistical methodology was applied later, including a comprehensive and transparent explanation on what assumptions and tests were made to account for capturing the attention and its relationship with the market, to gather both elements of investor sentiment.

5 Research design

In the following chapter, I will be focusing on developing and designing a well-grounded and refined modeling framework. To achieve this, I will be continuously considering all the relevant information gained from the theoretical overview of the biases and limitations in the previous chapter. This way, I want to ensure that my methodology always follows and remains in line with the lessons learned there.

5.1 Data overview

Before digging deeper and defining the cornerstones of the model framework, the datasets should be introduced in detail: the search volume downloaded from Google Trends and the to-be-predicted Dow Jones Industrial Average index. In the case of the Google Trends data, the main specialties and limitations (which would be handled later on) were already introduced earlier. In this section, I will be focusing more on the technical details. For the descriptive statistics of the datasets, the standardized search volume data of the different keywords, and the log return of the DJI, please refer to the Appendix to Tables 4 and 5 respectively.

5.1.1 The main characteristics of the Google Trends data

I intended to utilize Google Trends data in the best possible way, which means that search volumes are downloaded from the beginning of 2004 until the end of August of 2020 with a daily frequency. The first problem, one has to face, is the data download management. Based on the standardization method I chose (for details please refer to Section 5.2), I downloaded the dataset in 90 days long batches independently, with a one-day lag between each of them. Per keyword, this translates to slightly more than 6000 data requests, which

results almost in a 4.25 days long downloading time if one paces the requests one per every minute in order to avoid evoking the daily ban from Google Trends. I am aiming to use 30 keywords in my research (please refer to Section 5.3 for the specifics of keyword selection), which concludes ca. 125 days of pure data download, assuming that only one IP address is used. To slightly counterbalance this huge estimate, I rented 30 servers from Amazon Web Services, each having an own IP address and elastic computing capacity to be able to conduct the data downloads for one of the 30 keywords. So in the end I managed to cut back the time spent by search volume download to 4.25 days in total.

As the above details indicate, the collection and structuring of the Google Trends data is by far the most time and capacity consuming part of the whole process. For handling the querying of the Google Trends data I used the ‘gtrendsR’ package which provides a relatively simple toolset in the R programming language to download data from Google Trends. The functions of the library permit the user to launch a data request towards Google Trends by giving only the main parameters: the keyword, the time frame, the geographical location, and other technicalities (Massicotte, 2019).

Since I was planning on forecasting U.S. market movements, I had two valid options for what geographical location I wanted to examine the searches. I could either choose to concentrate on solely U.S. searches, or I could consider the search volume of the whole world. Preis et al. (2013) found that strategies based on global search volume data are less successful in anticipating movements of the U.S. market than strategies based on only U.S. search volume data. Later articles (including other fields of studies) tend to also rely on the focused search volume of specific countries when building their prediction models. Therefore I proceeded similarly, I also chose to use the U.S. search volume as the underlying data for prediction.

5.1.2 The target variable: the Dow Jones Industrial Average index

For forecasting the stock market trends of the U.S. a good proxy is needed. As most of the similar researches, I am going to use one of the most popular and fundamental U.S. based indices: the Dow Jones Industrial Average index. Returns are calculated as the logarithm of relative portfolio changes derived from the daily adjusted closing prices, following the usual definition of returns. As per my hypothesis, today’s search volume (v) should explain the price (P) change in the near future, thus the calculated log returns (y) in the near future. It would be quite obvious to apply that today’s search volume explains tomorrow’s log return, but as it was mentioned in Section 4.1, we should apply at least a three-day-long lag due to Google’s limitation. Therefore the formula which we are going to work with will look like this:

$$v_t \rightarrow \ln \frac{P_{t+3}}{P_{t+2}} = y_{t+3}$$

Obviously, there are fewer trading days than days when we have valid Google search data, and there were also a few cases where the opposite was true. I applied an inner join between the Google Trends or the DJI datasets in order to eliminate the missing or NA data points.

5.2 Standardization method

As it was mentioned in the previous chapters when listing the main limitations of the data, using daily data is a clear objective for us to be able to utilize Google Trends data in an efficient way. It was also noted that a careful standardization method is needed in order to gain comparable data points on the same scale. I strongly believe that this is an essential core step to have a usable, unbiased dataset to begin to build a modeling framework onto. What is very surprising and rather disturbing that none of the articles in this field of study (which are included in the literature view), not even the most popular and most cited ones, mention this step. It seems from those articles that after the data download they begin to work with an unstandardized raw dataset directly. In the following chapters, I will be calling attention to why the negligence of this step can lead to serious statistical issues which can endanger the later modeling process and render the results biased and completely incorrect.

Even though I have found no standardization solution (not even mentions of it) in the literature, one can get aid from informal sources such as acknowledged programming forums, data science blogs, and private web pages. One big issue here is that there are methods discussed in these sources which, although provide a standardized dataset, use a statistically incorrect method. I will be presenting one of the most popular standardization methods that I have met many times during my research, and I will show its shortcomings. In the end, I will be showing two approaches that supply a statistically unbiased dataset, and either one of these should definitely be applied before beginning to work with the Google search volume datasets.

In my examples and explanations, I will be assuming a “perfect world”, where the Google search volume of a given day is readily available at the end of the same day. This is only for the purpose of simplicity and transparency. Unfortunately in the “real world” one has to wait at least 36 hours for the Google service to summarize the most recent daily data, as mentioned in Section 4.1. When utilizing the Google data later, this characteristic will be also taken care of, and incorporated into my models, since the only thing that this delay changes is the lag that we have to apply in our models to begin the prediction process. To sum up from a technical perspective, in these examples there will be no extra lag applied (for presentation purposes), meanwhile, in the actual trading models (introduced in Section 5.4) the 36 hours will be added as an extra 2 days lag.

5.2.1 The dangers of working with unstandardized data

The dangers are rooted in the technicalities of how Google supplies the search volume data through the Google Trends engine. Therefore it could be difficult to see and grasp this for researchers who did not take a deep dive into the documentation of the data service. As I have described, Google Trends shares the relative search volume of a keyword in a given time span. The relative values are scaled between 0 and 100 according to the local maximum search value of the time frame: the local maximum gets 100, then the rest are scaled relative to this. Unfortunately, this scaling method hides two possible issues that should be handled in order to get comparable, standardized, and unbiased datasets.

The first problem is that the comparability between different time frames is not ensured this way. As an example, let me take a 180 day long period of search volume, that I want to use as a whole for prediction purposes in my model. As per the restrictions of Google, I can get this 180 daily data with at least two independent data downloads: one for the first 90 days, the second for the next 90 days. However, now we have two datasets at our disposal which are independently scaled on the 0-100 relative scale. Imagine an extreme case, where the absolute daily search figures (V) of the first 90 days are only the fragments of the absolute daily search figures of the second 90 days period respectively, thus $V_t = \delta * V_{t+90}$, where $t = (1, \dots, 90)$ and $\delta < 1$. In this case, the relative scaling happens independently upon the data download as well, and each of the time spans would be scaled separately, relatively to their local maximum values. Even though the absolute search volumes between periods did not match, after the relative scaling we would get exactly the same relative search volumes (v) for each of the separate periods (if the above equation stands), thus $v_t = v_{t+90}$, where $t = (1, \dots, 90)$. It would be an utter mistake to just simply bind the two time series together, and use it as a 180 days long period, because as the example indicated the absolute values are on a totally different order of magnitude between the two periods. Only the relative scaling makes it look like the two time spans are comparable without any adjustments to the data. We can see this through an example in Figure 1 below.

day (t)	absolute search volume		relative search volume		$\delta = 5\%$
	$V1_t$	$V2_t$	$v1_t$	$v2_t$	
1	15,000		30		
2	10,000		20		
(...)	(...)		(...)		
37	50,000		100		← local maximum
(...)	(...)		(...)		
90	35,000		70		
91		300,000		30	
92		200,000		20	
(...)		(...)		(...)	
127		1,000,000		100	← local maximum
(...)		(...)		(...)	
180		700,000		70	
$V1_t = V2_t * \delta$			$v1_t = v2_t$		

Figure 1: First problem of working with unstandardized data

Source: Own figure

The second problem is even more subtle than the previous one. It is actually only a problem if one wants to use the daily data for daily forecasting, which is exactly the purpose of Google data in this entire field of study. The problem stems from the look-ahead bias, let me demonstrate it through another extreme example (presented in Figure 2). Imagine that in the present we want to build a daily prediction model with historical Google data. We could use a period of 90 days (or less than 90 days), this way we could avoid the first problem of linking together multiple time frames. Upon downloading the data, we face that the local maximum is on the 90th day. This makes it clear that we are not allowed to use the information of these 90 days to predict anything within this time frame, only for the days coming after. The reason is again the relative scaling. We can't use the data point of day t because it was scaled based on information that came into light only on the last day of our time frame. The relative data on day t would not be the same without knowing future information, in our case the local maximum search volume of the given time frame, which was then used in the scaling process of the entire downloaded data set. A local maximum (or any type of local) value of a time series can only be named when the entire dataset is revealed in time.

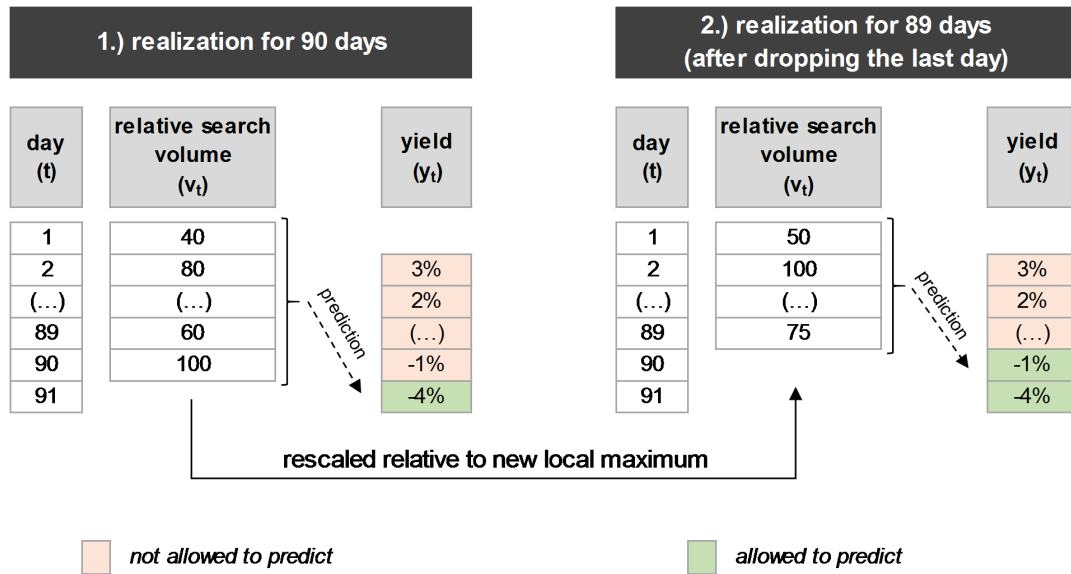


Figure 2: Second problem of working with unstandardized data

Source: Own figure

Furthermore, it can be proved, that this is not only a problem if the local maximum is on the last day of our downloaded time frame. Since we are talking about a local value, it is highly dependent on the choosing of the time and length of our time frame. Circling back to our case, we could just simply say that we are going to take another 90 day period with a one-day lag behind the original one or just an 89 day period without the 90th day, making the previous local maximum drop from our time frame. However, now every data point is rescaled based on the new local maximum of the new time frame. We can see that theoretically, it could be possible to use certain data points within a given time frame for forecasting within that time span, for example, the global minimum or maximum point. But to find the global extreme values of the whole time frame we intend to use, we would have to try every possible combination of time frame placement and length to make sure we have the correct value (which, goes without saying, would be a waste of time and resource). It is much more convenient to adhere to the principal not to use any of the search volume data of a given time frame to forecast any external data within that time frame, to avoid any effects of the look-ahead bias.

We should also note that the second issue can be treated in a simple way: a comparison between data points should remain relative, and it must not become absolute. Continuing with the above example in Figure 2, we can see that the relations and the ratios between data points remained the same, however, due to the rescaling the absolute differences are rescaled too. Therefore, if we can stick to this principle, we can always dodge the inherent look-ahead bias. Additionally, there are standardization methods that could be utilized to deal with the above two issues, that I would be presenting below. Unfortunately, all of these approaches are sourced from informal sources and further developed by me, which

puts the question forward whether any of the above-described issues were recognized, and even treated, previously in the literature at all.

5.2.2 The two-step weighing method

During my search for solutions to standardize the Google data, mostly I encountered the two-step weighing method. This method is popular because it is transparent, not too difficult to implement, and requires only a minimal number of data download requests (which is crucial due to the limitations of Google Trends). However, this method only solves the first problem I stated above, and not only does it not give an answer to the second problem in connection with look-ahead bias, but it further amplifies it. I will touch upon that after describing the details of the approach in depth.

Month (m)	Monthly data (w_m)	Day (t)	Daily data ($v_{t,m}$)	Standardized daily data ($v^*_{t,m}$)
2004 - January	53	2004-01-01	40	$40 * 53 / 100 = 21.2$
		2004-01-02	50	$50 * 53 / 100 = 26.5$
		(...)	(...)	(...)
		2004-01-31	55	$55 * 53 / 100 = 29.2$
2004 - February	30	2004-02-01	100	$100 * 30 / 100 = 30.0$
		2004-02-02	44	$44 * 30 / 100 = 13.2$
		(...)	(...)	(...)
		2004-02-28	23	$23 * 30 / 100 = 6.9$
(...)	(...)	(...)	(...)	(...)
2020 - Aug	65	2020-08-01	67	$67 * 65 / 100 = 43.6$
		2020-08-02	100	$100 * 65 / 100 = 65.0$
		(...)	(...)	(...)
		2020-08-28	56	$56 * 65 / 100 = 36.4$

one data request
seperate data requests per each month

Figure 3: Two-step weighing standardization method

Source: Own version based on the figure of Bewerunge (2018)

As Figure 3 shows, the two-step standardization exploits that Google Trends provides daily data within monthly intervals and monthly data within intervals above five years. First, daily (t) data is obtained for every month (m) separately and independently. As a result, the user is given daily datasets ($v_{t,m}$) of separate months, scaled entirely independently based on the local maximum of the given month. Secondly, with one query, the monthly data points (w_m) are requested for the whole interval (which should be above 5 years to get monthly periodicity). These monthly data points are also relative figures, but they are scaled in the same range since they were provided with one data request. After that, the monthly data points will be basically acting as weights in the standardization process, and they get assigned to the corresponding daily data points: one monthly data point becomes

the weight of the entire daily dataset of that given month. By using this method the standardized daily data points ($v_{t,m}^*$) are now comparable spanning over months through the whole timeline (Franz, 2018).

$$v_{t,m}^* = \frac{v_{t,m} * w_m}{100} , \text{ where } t \subset m$$

This method takes care therefore the first problem, it scales independent datasets onto the same scale, making the data points comparable across the whole time frame. Yet, it completely fails the principle of the second problem. On one hand, we introduce a look-ahead bias when using the daily data points within the months to predict within that same month, because those were scaled relative to that given month's local maximum value (which would not have been available information without knowing the daily dataset of the entire month). On the other hand, we are getting another look-ahead bias involved with the monthly data points, considering that the same guidelines (regarding relative scaling) should be adhered there as by the daily data points. Furthermore, this second look-ahead bias presumably has an even larger order of magnitude, because it compromises a period which is at least 5 years long.

Despite all these facts above, it should be noted that this approach is perfectly appropriate for standardizing Google Trends datasets. In case of a descriptive statistical analysis about search volume or for the purpose of visualizations it can be a transparent and straightforward approach to gain a standardized, comparable time-series dataset. We can even use it in a prediction model too if we only observe changes relatively between data points. However, the monthly data points are not available until the end of the given month, thus, they can't be used as weights in a real-time prediction scenario. This makes this method rather impractical, especially since we are trying to create a model that can be utilized for actual trading.



5.2.3 The rolling time frame method

To come up with a method that is usable for forecasting, we have to go back to the two fundamental requirements that we stated above (the first and the second problems). To avoid the second issue, the inherent look-ahead bias due to the scaling, we only have to ensure that the search volume downloaded for a given time frame can only be used for prediction from the time when the search volume of the last day was revealed. This can be achieved with backward-looking statistical methods, such as using the dataset in regression models, for momentum calculation, etc. So the information gained from the search volume downloaded on a given day is only used to predict market yields of the next day at the earliest.

Now, that the second problem should be taken care of, we can concentrate on the first

problem. Here the issue is that we want to bind together search volume datasets coming from independent data downloads. But is binding together independent search volume datasets an absolute necessity to create a prediction model? Imagine that with the search volume on day t we want to predict the market yields of day $t + 1$, thus $v_t \rightarrow y_{t+1}$. For this, we can download the search volume of the last days (up to 90), and build an algorithm to gain information from these few days of the historical dataset, and make the prediction with it. On the following day, on day $t + 1$, we can then repeat this process for forecasting the market yield of day $t + 2$. We just have to move (or even expand if we did not reach the 90 days length limit previously) the previous time frame one day ahead, repeat the data download, then gain the necessary information from this dataset. Meanwhile, we are absolutely neglecting the previously downloaded dataset, thus there is no need for handling those together. If we want to extend our time frame, this will result in a rolling prediction model, since the historical time frame (used for training) always will be rolled one day ahead, as Figure 4 shows.

day (t)	yield (y _t)	relative search volume		
		v1 _t	v2 _t	v3 _t
1	3%	100		
2	2%	40	30	
3	-4%	70	55	80
(...)	(...)	(...)	(...)	(...)
89	-1%	20	18	23
90	2%	60	49	68
91	-2%		25	30
92	0%			42
93	3%			

 not allowed to predict
 allowed to predict

← - - - ' predictions

Figure 4: Rolling time frame method

Source: Own figure

With this rolling time frame method, we are not risking any chance of involving a look-ahead bias, and we managed to avoid binding independent time frames together, thus we overcame the second and the first problem (respectively). One could argue that this way our modeling framework is highly restricted, since we can only consider and rely on maximum of 90 days of historical data for each prediction we make, and we cannot look further back from that. This method also prevents us from being able to compare independently downloaded search volume datasets, since those are scaled on different spectrums, according to their local maximum values.

5.2.4 The overlapping time frame method

In the previous section, I have shown how we could bypass binding together independently downloaded datasets, while building a rolling model to even forecast on longer time periods, without having to worry about introducing look-ahead bias. But what if we would like to utilize the whole historical dataset, and build an expanding model? In this case, we would not be able to simply dodge the first problem, we are forced to join datasets together to cover the whole time frame. So far, I have only found one standardization method, which fulfills the requirements of both stated issues, and creates a standardized overall dataset for time spans of any length, while avoiding any type of look-ahead biases totally.

I call this method the overlapping time frame method because the key of the standardization lies in the overlapping part of two (or more) independently downloaded search volume time series. If two individually downloaded Google Trends datasets on different time intervals (v_1 and v_2) have an overlapping part, then we can calculate an average ratio ($A(h)_{v_1, v_2}$ where h is the length of the overlapping part). This ratio is calculated as the average of the individual ratios (α_t) of the data points of the two intervals within the overlapping part. With this average ratio, acting as a weight, we are able to standardize the two datasets onto the same relative scale to make them comparable, by simply multiplying (or dividing) the values of one of the datasets, as shown in Figure 5 and the formula below. Please note that this method can be extended, and it lets us chain together multiple datasets. It is also important to highlight, that with this method, we might leave the 0-100 scale during the rescaling process (by exceeding 100), but that should be expected if we want to correctly relate and compare two (or more) independent datasets.

$$A(h)_{v_1, v_2} = \frac{\sum_{i=1}^h \alpha_{t+i}}{h} = \frac{\sum_{i=1}^h \frac{v_{1,t+i}}{v_{2,t+i}}}{h}$$

Since we are using an average ratio to standardize data points, there are some technical specialties that should receive extra attention, in order to minimize the errors coming from the averaging. First, one can easily notice that we can get rid of the average function, when the overlapping part is minimal, thus, it is only one day long ($h = 1$). In this case, the overall ratio will be equal to the sole individual ratio: $A(1)_{v_1, v_2} = \alpha_{t+1} = \frac{v_{1,t+1}}{v_{2,t+1}}$. This would be an optimal solution if the independently downloaded search volume realizations were always identical for the same time frame. Unfortunately, this is not the case, and it is because of the sampling extraction method of Google Trends, presented in Section 4.1.4. Due to the sampling extraction method, the daily search volumes might fluctuate.

To combat the fluctuation and to maximize the precision of the average ratio ($A(h)_{v_1, v_2}$), we should calculate it using the most possible data points, thus we should maximize the length of the overlapping period. Therefore, in case we are binding together two time frames of the same length of h ($h(v_1) = h(v_2) = h$), we would like the overlapping part

to be the length of $h - 1$. In the rare cases when the length of the time frames don't match ($h(v1) < h(v2)$), we want to take the length of the shorter time frame ($h(v1)$) and adjust the length of the overlapping time frame accordingly (to a length of $h(v1) - 1$). Secondly, it is recommended to always standardize the later dataset onto the scale of the earlier one. In addition, for the overlapping part, one would want to keep the data points of the earlier dataset, which are yet untouched. By doing so, only the data points of the later dataset, which are outside the overlap, should be modified by the average ratio. By following the above-described principles, we would be minimizing the error coming from the precision of the calculated ratio, because at each standardization step (meaning the binding of two independent time frames) we would be standardizing only one data point: the last data point of the later dataset, the one that is outside the overlap. Please see examples in Figure 5 below.

example for $h = 1$					example for $h = 89$				
day (t)	relative search volume		standardized search volume (v^*_t)		day (t)	relative search volume		standardized search volume (v^*_t)	
	$v1_t$	$v2_t$				$v1_t$	$v2_t$		
1	100		100	x_1	1	100		100	x_1
2	40		40	x_2	2	40	21	40	x_2
(...)	(...)		(...)	(...)	(...)	(...)	(...)	(...)	(...)
89	20		20	x_{89}	89	20	9	20	x_{89}
90	60	30	60	$x_{90} = A * y_{90}$	90	65	33	65	$x_{90} = A * y_{90}$
91		55	110	$A * y_{91}$	91		70	140	$A * y_{91}$
(...)	(...)	(...)	(...)	(...)					
178		100	200	$A * y_{178}$					
179		80	160	$A * y_{179}$					

$h = 1$

$A(1) = 2$

$h = 89$

$A(89) = 2$

Figure 5: Overlapping time frame method

Source: Own figure

5.2.5 The chosen standardization method

To totally avoid any look-ahead biases while creating a strategy usable in actual real-time trading, only the last two approaches can be selected as viable standardization methods. The trade-off between the two comes from different aspects. On one hand, by using the rolling time-frame method, one would give up the chance to use historical search volume data (beyond the 90 days period) to train a prediction model. On the other hand, by choosing the overlapping time frame method, unintentionally a small standardization error might be added to the scaling due to the average ratio. My research plan consists of validating and testing methods of earlier researches, so I am bound to use the entire dataset, to be able to run not only rolling but expanding statistical training algorithms as well. For this reason, I decided to standardize my downloaded search volume dataset

with the overlapping time frame method. In my particular case, by keeping the above principles in mind, that meant that for each of my keywords I had to download search volume data sets of 90 days spans with a one day lag beginning from 2004 until the end of August of 2020. This justifies the huge data and computing capacity necessity described in the earlier sections about the characteristics and limitations of the data.

The above arguments prove the obvious need for a standardization method in order to get a comparable dataset that is standardized onto the same scale and free of any look-ahead biases. For this reason, I would like to once again highlight that I was very surprised that none of the literature (listed throughout my research paper) mentions this inevitable step in their methodologies. This is despite the fact that most of them explicitly declare the use of expanding training methods, or published charts and figures displaying search volume indices spanning over many years (or even almost decades). However, all these would be statistically incorrect without a proper standardization step involved first. Hopefully, not mentioning the standardization procedure is only an oversight, because the actual lack of it in a model might render all its results and outputs unreliable. Nonetheless, one should always check for standardization methodology in a paper using long time frames of search volume data that could not have been produced from one single data download. If in this case no standardization steps are introduced, I would suggest interpreting the results and forming serious conclusions based on them, very cautiously.

5.3 Choice of keywords

When selecting the keywords for my study, I tried to dodge every issue that I listed before. In this chapter, I will present the selection procedure, with all the reasoning included to ensure an objective and unbiased choice. I especially kept an eye on avoiding any possible look-ahead and survivorship biases.

5.3.1 Arguments for choosing the top-down approach

I have decided to use a top-down approach in the procedure of selecting keywords for the following several reasons. First, in my research, I definitely wanted to focus on capturing sentiment, since I believe that attention and its relationship to the market should be examined and handled equally importantly in order to build a comprehensive framework. In my opinion, the top-down approach (using generic and economically relevant terms) is more appropriate for this purpose, because the bottom-up approach (using company names, tickers, and industry-specific terms, etc.) is usually paired with concentrating more on investor attention in the trading model. Secondly, to build up a good representation of the DJI from the search volume of company-related keywords through a bottom-up approach, we would have needed a vast amount of keywords, thus lots of data requests. Due to the data request limitations of Google, this is highly impossible without serious compromises,

which could indeed cast a shadow on the quality and reliability of the research.

Thirdly, the arguments of Da et al. (2011), concerning choosing company names as keywords, are very discouraging. Going with tickers is less risky but we can also have problems with the ‘noisiness’ there. My goal is to develop a model that captures the sentiment of the public, therefore capturing mostly trade-related searches via tickers is also not in our best interest. Last but not least, two of the most popular papers of this area, written by Preis et al. (2013) and Da et al. (2015), also used a top-down approach, so it seems practical to do the same. However, now that the top-down approach was decided, we also have some key points to consider in terms of what keywords we should include specifically.

5.3.2 The finance-related keywords

First, we should choose finance-related keywords that have either a negative or positive economic sentiment inherently, and (what we can derive from the former studies) we should choose these based on an acknowledged source. The fact that our finance-related terms will have a clear sentiment gives some practical perks too because this sentiment could determine the relationship of the keyword with the market movements. In the case of a keyword with a negative economic sentiment, a higher search volume would mean panic on the market, while a lower one would mean a calm market or even euphoria (and in the case of a positive one vice versa).

The WordStat Sentiment Dictionary (created by Loughran and McDonald in 2011) can be considered as a perfect source to satisfy our needs. In the describing article, the authors provide a clear demonstration that most of the negative and positive words in the general Harvard IV TagNeg dictionary are typically not negative or positive (respectively) in a financial context. Therefore, they created a custom dictionary specifically for the financial niche (which was published later by the Provalis Research Institution), listing words that have negative or positive economic meaning.

Eventually, I chose 10 negative and 10 positive keywords from this dictionary to become the subjects of my prediction model. The widely cited flagship keyword of Preis et al. (2013), “debt” can be also found in this dictionary, among the words with a negative economic sentiment. By including it in our keyword selection gives an extra benchmarking opportunity here.

5.3.3 The control group

Last but not least, the warnings of Challet and Ayed (2013) ought to be seriously considered as well. Their main point is to have a control group of keywords, that are not biased by finance, for benchmarking purposes. Therefore, I am going to add 10 keywords as a control group to the finance-related ones. Choosing keywords for the control group was performed following my own consideration. I wanted to find as general words as possible that will

have a sufficient amount of search volume. With these, I wanted to cover a wide range, so in addition to general words I added e.g. obviously seasonal keywords (“christmas” and “swim”), a geographical name (“chicago”) and even an often misspelled word (“bycicle”). To not have any finance-related terms in the control group by accident, I checked them against the WordStat Sentiment Dictionary of Lougrand and McDonald (2011). By avoiding keywords presented there, not only means that the keywords of the control group are not finance-related but that they don’t even have any economic sentiments (they are neutrals). The selected financial keywords with positive or negative sentiment and the neutral keywords (the control group) are shown in Table 2 below.

financial keywords		control group
<i>positive (+)</i>	<i>negative (-)</i>	<i>neutral (o)</i>
reforms	debt	banana
boost	crisis	door
consolidate	decline	weather
construction	recession	chopstick
outperform	unemployment	phone
savings	bankrupt	chicago
progress	deficit	christmas
booming	collapse	swim
accrue	market crash	bycicle
surpass	downturn	programming

Table 2: Keywords used in the model

Source: Own table

5.4 Development of the trading strategy

In this chapter, I am going to set the course of the development of a trading strategy, which is intended to be statistically unbiased and fully operational for daily trading activities. I will begin the process by taking the strategy of Preis et al. (2013) as a starting point, then I will introduce every type of modification and advancement to it step-by-step. At each stage, following a thorough explanation, my hypothesis will be stated, but the test results will be presented in the next chapter. It is important to note that the same approach was taken not only here, at the planning phase, but at the execution part too, in order to avoid any doubts of p-hacking. To avoid any type of statistical biases, I will be conducting the tests by strictly adhering to the principles and methods I highlighted in the previous chapters (such as utilizing correctly standardized data, using the control group for my positive and negative keywords, applying an expanding training and testing method, etc.).

5.4.1 The “literature-based” approach: recreating the method of Preis et al.

As I have already highlighted, the research paper of Preis et al. (2013) has been a highly influential and popular article in the field of Google search prediction. I have decided to use it as a starting point for several reasons. First, its modeling framework is not too complex despite achieving prominent results. Secondly, the methodology described in the paper is unfortunately rather laconic on certain key decisions and on the backgrounds of hypotheses, so there is room for further testing and documentation. Thirdly, it will give an excellent benchmarking opportunity to compare not only the results but the fundamentals of the frameworks as well.

Preis et al. (2013) used weekly search volume data of 98 keywords to determine the price change of the Dow Jones Industrial Index from 2004 to 2011 (prices were obtained for every first trading day of the weeks). They compared the search volume of week t (u_t) to the average search volume of the past $1, \dots, 6$ weeks (Δt). If the relative change of the search volume was positive, then they took up a short position, and if it was negative, they entered into a long position until the following week when a similar decision had to be made. The strategy is reflected by the below formula.

$$position = \begin{cases} long, & \text{if } \Delta u_{t,\Delta t} \leq 0 \\ short, & \text{if } \Delta u_{t,\Delta t} > 0 \end{cases} \quad \text{where} \quad \Delta u_{t,\Delta t} = \frac{u_{t-1} + \dots + u_{t-\Delta t}}{\Delta t}$$

They reported the average yields (over $1, \dots, 6$ weeks) in the end, and other than that, they often highlighted the striking results of the keyword “debt” with a 3 weeks average ($\Delta t = 3$) throughout the paper in many visualizations and tests. I haven’t found any written statistical proof on why they settled on using the average of the past 3 weeks for this purpose, other than the fact that it resulted in the highest yield. According to my hypothesis, there will be no significant differences between using different weeks, and that is probably the reason why Preis et al. (2013) did not highlight this decision in their research. Based on these, my first task is to assess the difference between using the moving average of different weeks (1 to 6) by reconstructing the described framework of the article. Since I have three keywords that match theirs (“debt”, “crisis” and “unemployment”), I will be assessing the results of these three on the same time horizon that they have used (from 02/01/2004 to 22/02/2011), to be able to compare the performance of our models directly. For later benchmarking purposes and more proof of concept, I will also perform the same tests for all of my 30 keywords for the originally appointed testing period (ending on 28/08/2020). I also applied one more change, instead of using search volumes with a weekly periodicity, I used daily datasets. In my case, their weekly data points translate to more frequent daily data points, thus, the moving average of the past $1, \dots, 6$ weeks was replaced by the moving average of the past $7, \dots, 42$ days. Regardless of the test results, I

will be continuing my further analysis using the average over the past 3 weeks, since the authors reported those results as their main achievement.

5.4.2 The “dictionary-based” approach: introducing dictionary-based sentiment

There is one further unexplained assumption of the above-described strategy of Preis et al. (2013). They are making a deterministic relationship between investor attention and sentiment because they are assuming that no matter the characteristics of a keyword, increased attention determines a negative sentiment, thus, an increase in search volume compared to the average of the past weeks results in price falls of the market, and vice versa. However, there is absolutely no reported evidence in their article that would justify this decision.

My hypothesis is that this relationship does not stand, the investor attention not directly determines the investor sentiment, and the evaluation of the relationship should be needed. I will be testing this on my previously presented keyword pool (10 positive and 10 negative keywords based on the Loughran-McDonald (2011) WordStat Financial Sentiment dictionary, and the control group consisting of 10 neutral words). I assume that the deterministic relationship presumed by Preis et al. (2013) might be only correct for the keywords with a negative sentiment. In this case, the increased search volume of a negative keyword would imply an increasing negative sentiment on the market, which could lead to taking up a short position. While a decreasing search volume for that given negative keyword would mean that the market outlook is getting more positive, so we can enter into a long position. The opposite stands for positive keywords: an increasing search volume can be translated as a growing positive sentiment, therefore, a long position can follow, and vice versa. If this assumption stands, then the search volume of the neutral keywords should have no relationship with the market movements. Therefore, if I test them both as positive and negative keywords, they should have perfectly contrary but not significantly good results when compared to the positive and negative keywords.

5.4.3 The “data-based” approach: quantifying the relationship

In order to conduct an extensive investigation, I am going to introduce a method to determine the sentiment of my keywords by a data-driven method. For this purpose, I am going to use expanding rolling univariate linear regressions. There are several reasons why I decided to take this approach.

Kogan et al. (2009) used a similar regression-based method to predict volatility by running a text-analysis software on the annual reports of the publicly traded companies (required

by the SEC²) over the period between 1996 and 2006. They have found that this is not only an objective way to measure the relevance of a certain keyword, but it also accounts for the fact that the relevance of keywords is not constant over time. Da et al. (2015) took a similar way to construct their FEARS index. They built it up by search volume of Google keywords having negative significant predictive power, and they tested it against returns, volatility, and fund flow figures of the U.S. market from 2004 to 2011. They used an expanding linear regression every half a year to predict the direction and the strength of the relationship of their keywords with the market (hence my decision on the expanding modeling framework), accounting for the second element of sentiment prediction. Then they constructed their FEARS index according to the t-statistics of each keyword, thus, the significance of each. They even mention examples where they got contradictory results, in terms of the keywords' sentiment, to the Harvard sentiment dictionary. They emphasize that if they were to follow the lead of the dictionary, these findings would not have come to light, and their model framework would have been biased and misled.

In my model, I am going to combine the expanding linear regression approach of Da et al. (2015) and the moving (rolling) average method of Preis et al. (2013). Da et al. (2015) only ran their regressions once every half a year, thus they would act on one result for half a year when they rerun the expanding regression again and continue this cycle. However, this method assumes that the sentiment of a word, and more specifically the strength of this sentiment with the market, would be constant over this half a year. Avoiding any incorrect biases introduced by this assumption, this sequence could be made more dynamic. Since we have daily data, this sequence could be rerun on a daily basis for each keyword separately, after being able to acquire the recent daily search data. This way we acknowledge that the sentiment can change even over days, and still if we experience that the direction and strength of the sentiment remained relatively constant over periods of 6 months, we arrive at the same model as Da et al. (2015) did. We will begin the trading activities from 19/09/2004 to ensure that enough data points (180 records) are provided to meaningfully complete even the first linear regression. To prepare my search volume datasets for the regression, I will be taking the logarithm of the data first, then differentiating it once (in other words taking the log change of the data), similarly to the referred article.

In my model, the expanding linear regression will be responsible to predict the direction and the relevance (the strength) of the sentiment of a keyword, while the relative change compared to the moving average will capture the change in investor attention. The regression will first evaluate the strength of the linear association between the historical adjusted search volume and the historical yields on 95% significance level, thus with the critical values of ± 1.96 of the two-tailed t-distribution. Then if the regression is historically significant, the direction of the relationship would be determined: if the beta is positive,

2. Securities Exchange Commission of the United States of America

then we can conclude a strong positive relationship of the keyword and the market, and in case of negative beta vice versa. If the regression is not significant, we are not taking up any positions for that given day. Finally, we can predict what position we should take up by applying the rolling average part (focusing on entirely capturing the attention). This way the observed relationship and attention changes shape the sentiment through an entirely data-driven approach, which is in line with the earlier definition of these concepts. It is necessary to note the importance of the “data-based” sentiment evaluation method in comparison with the other two approaches. The difference lies in how the relationship is defined, how attention translates to sentiment. In the “literature-based” approach, Preis et al. (2013) simply assumed that this relationship is always negative and strong, hence, they take up a short position every time when they observe an increasing search volume and a long position in case of a decreasing search volume. In the “dictionary-based” approach, we tried to overcome this and introduce polarity (in the case of the sentiment) to the model. There we rely on the dictionary to assess the direction of the sentiment of a keyword, and we take this as a constantly strong relationship with the market. Finally, the “data-based” approach entrusts the whole evaluation process of strength and direction onto the data itself, ensuring complete objectivity, and acknowledging that these characteristics can and might change over time. For the above reasons, I strongly believe that the latter (or a similar data-driven) approach should be always applied.

To gain a deeper understating of the characteristics of the relationship of certain keywords and the market, first I will investigate and provide descriptive statistics based on the parameters of the regressions. Only after that will I move onto evaluating the performance of the trading model. Unfortunately, the performance of this approach might not be correctly compared to the previous two methods because there is a huge difference in the approaches. While we are taking up either long or short positions in the “literature-based” and “dictionary-based” approaches, we introduced a (third) null position in the “data-based” approach because there we are not taking up any positions if the relationship is not significant enough. If we assume that these relationships are not constant over time, we have to also assume that there might be long periods of time when we will be sitting in this null position, when none of our keywords are triggering a position. Since we are only considering one signal (the search volume of a certain keyword) this can mean that we will miss out on most of the happenings of the market, and only trade on certain events. This also foreshadows that while the “literature-based” and “dictionary-based” approaches are rather comparable to long-term trading strategies (such as a buy-and-hold strategy), the described “data-based” approach will rather be comparable with short-term benchmarks. So while the latter may result in a similar performance as the former two strategies, it is advised not to establish any conclusions based on solely the performance comparison. However, to make the “data-based” method a comparable trading strategy in

terms of performance as well, I will be introducing a new aspect to it in the next section. Meanwhile, I will try to focus on the analysis of the direction and strength of the assessed sentiment in this exercise.

5.4.4 The “data-based downside” approach: trading only on negative sentiment

In order to tune the “data-based” approach, the null position should be eliminated or replaced. Tetlock (2007) conducted a study where he systematically explored the characteristics of the interactions between the content of written media (the Wall Street Journal) and stock market activity (represented by the DJI) over a 16-year period 1984–1999. He was categorizing the most used words into 77 different categories, such as negative, positive, weak, passive, pleasure, etc., via a text analysis software (General Inquirer program) to create his input datasets, and he also created a pessimism index based on the linear combination of said categories. One of the key findings of Tetlock (2007) was that high values of pessimism induce a decrease in market prices because presumably, it affects the behavior of small individual investors. His principal component analysis of the pessimism index also shows that this same relationship could be interpreted by examining only the effect of negative and weak words, which account for most of the variance (57%) in the first factor. He also clearly states that the effect of the positive sentiment (observed through the index and the categories themselves) is not as strong as the effect of the negative sentiment, the words that were deemed to have a negative sentiment contributed the most to map out the relationship with the market. Da et al. (2015) confirm these findings: overall from their 118 terms they found none that had a t-statistic of at least +2.5 (indicating a strong positive relationship), however, they have found 14 keywords that have a -2.5 or lower t-statistic (meaning a significant negative relationship).

Based on these facts, I am proposing a slight modification to the logic of the “data-based” approach. For each keyword, I am going to only consider the downside signals (hence the name of the strategy) when the observed sentiment indicates possible price falls. By acting solely on the negative sentiment, only short positions are triggered by the “data-based” approach, thus, I will be only actively trading the downside. Otherwise, when the combination of the regression and the moving average does not trigger a short position (on the 95% significance level), I will be taking up a long position in the Dow Jones Industrial Index as a default position. By having the long position as a fallback, I completely removed the null position from the model. This ensures that the previously stated concerns do not apply anymore, therefore the performance of this model can and should be benchmarked not only to the buy-and-hold strategy but to the “literature-based” and “dictionary-based” approaches as well. However, no matter the performance, this approach is still the most dynamic and objective among the listed ones, so I will be using this in the later stages of

the model-building.

I feel also obliged to test whether the assumption of Tetlock (2007) and later Da et al. (2015), meaning the reasoning behind only using the negative sentiment for a trading signal, can be confirmed via my sample as well. Therefore, I will be taking a step back, and perform backtesting whether the yield of the previous “data-based” approach came from rather acting on the positive or negative sentiments, so from the taken-up long or short positions (respectively).

5.4.5 The “data-based downside index” model: compiling the individual signals into one

The ultimate goal of Da et al. (2015) was to construct a list of keywords, and use their collective sentiment to reveal market movements. Technically, they achieved this by compiling an index out of the information they gained on an individual term level. Out of their 118 terms, they have used always 30, which had the most negative t-statistics, to trade accordingly. The list of 30 was recalculated every half a year, upon rerunning their expanding regression model. The forming of a collective index is needed, on one hand, to diversify the exposure to individual terms, thus, to reduce the chance of idiosyncratic noise, and on the other hand, this way, one trading decision can be backed by not only one but an aggregation of several strong individual signals.

However, having a word list with a fixed number of elements to fill at each rerun can be rather dangerous. Imagine if we did not get any keywords with a strong negative relationship towards the market for a cycle, then in this case we would be acting on a collection of words that has low or no relationship to the market at all. Presumably, in this described case, our strategy would perform similarly to a totally random trading strategy. From this example, it seems that the success of a similar approach eventually depends on the original pool of our keywords. In order to reduce the exposure to our keyword selection method, we would have to come up with a more dynamic approach that only considers the signals of the keywords that have a significant negative relationship with the DJI. Therefore, I am going to compile the signals of my 30 individual keywords, and even if one triggers the short position threshold, I am going to act on it. Please note that for a bigger sample of keywords the implementation of other restrictions and model constraints might be wise (such as only acting on a signal if at least a certain portion of our keywords breached the threshold), since the chance of taking up a short position increases with the number of our keywords at hand.

There is another key decision point in connection with the selection of keywords that can be tested via this method. Since the “data-based” approach was introduced, the “dictionary-based” characteristics of a keyword were neglected. This is not an actual issue because we are reassessing the sentiment based on the data itself on a daily basis. However, there is

another important aspect of the Loughran and McDonald (2011) dictionary: it collects the terms that have economic sentiment, thus, no matter whether positive or negative, the meaning of the keywords have an economic significance. Therefore, we can test whether words directly selected from the dictionary perform significantly better than the control group consisting of neutral keywords (which should not have an economic relevance). To have a robust test, I am going to first assess the performance of a “data-based downside index” consisting of solely positive, then negative, then neutral keywords. Then, I will be combining these groups: the positive and negative, the positive and neutral, and the negative and neutral words. Finally, I will combine all three groups of words into one pool, and repeat the exercise with it. Afterward, the performance of all the 7 different groupings could be compared.

My assumption is that a trading strategy based on solely the neutral keywords is rather random than economically justifiable, thus the keywords that have an economic significance (the positive and negative keywords of the dictionary) will outperform on an individual group level the neutral ones. Therefore, I believe that the combined group of the positive and negative keywords will surpass either of the groups which have neutral keywords in it, even the one where all three groups are combined. Since the evaluation of the direction of the sentiment is not dependent on the dictionary in the “data-based” approach anymore, all the keywords included in the dictionary could result in a strong negative correlation with the market. This means that the chance of finding a working signal might be bigger if we take both the positive and negative (determined by the dictionary) into our pool, consequently, the combined group of these words (basically the ones that have an economic sentiment according to the dictionary) should perform better than either of the sole groups of these terms.

Forming the individual signals into a collective one raises a rather technical but important question regarding the confidence level. This question arises because when multiple simultaneous hypotheses are tested at a set confidence level, there is an increased risk of committing Type I errors. So far all the univariate regressions were tested on a 95% significance level separately. However, our main model is now going to consist of 20 keywords with economic significance (not to mention the other tests for other groups of 10, 20, and 30 keywords as well), thus 20 different univariate linear regressions that can each give us a signal. With a 95% significance ($\alpha=5\%$) level we basically act on a model which 1 out of 20 times on average might signal us incorrectly, thus, one keyword out of the pool might just always signal randomly. Following the counsel of Feise (2002), we now have two paths ahead of us.

Classicists believe that in case of a situation like this a p-value adjustment should be applied, making the conditions of accepting a hypothesis more severe. The most acknowledged method to do so is the Bonferroni correction which adjusts the α linearly. In this case, our

individual regressions should be tested with the original α divided by the number of tests (n), thus $\alpha^* = \alpha/n$. However, opponents of p-value adjustments, the rationalists, have numerous objections against methods like these. One objection is that the significance of each collective test would be decided according to how many outcomes measures our collective hypothesis originally had. The decision of including more or less of these is still in the hands of the researcher, thus, this still allows p-hacking. Another argument is that if we decrease the chance of Type I errors, the chance of accepting false positives, we simultaneously increase the risk of Type II errors, the chance of finding false negatives. Practically, in our case making a p-value adjustment would mean that we decreased the chance of acting on incorrect individual signals, but at the same time, we have more risk of missing out on shorting opportunities due to the severe restriction of accepting the null hypothesis. Rationalists generally agree that p-value adjustment should be introduced based on the researcher's perception of the problem and not based on a generalized formula, involving a clear communication of why the adjustment was deemed necessary. Additionally to these two approaches, one could argue in favor of introducing one multivariate regression instead of the family of univariate regressions, but that also has its own disadvantages, such as interpretation problems.

Based on these guidelines I have decided to continue my testing with both of the two different approaches. Following the direction of the classicists, I will perform a p-value adjustment with the Bonferroni formula, increasing the confidence level in a very prudent way. At the same time, I will be also testing every case on a 98% significance level ($\alpha = 2\%$), as per the rationalist direction. In the case of the latter one, I have chosen to raise the significance to this specific higher value from 95%, because I wanted to avoid cases when I might be acting on just random signals of one of the twenty (or even thirty) keywords simply due to the low confidence level. I also did not want to be too severe, since I think the missed opportunities (Type II errors) are equally important to avoid acting on accidental signals (Type I errors). The critical values are always calculated following the two-tailed t-distribution.

5.4.6 Further tests regarding the strength of the collective signal

As a final test, I am going to investigate the relationship between signal strength and the actual length of holding a short position. For this purpose, I will use the positive and negative keywords, altogether a pool of 20 keywords, and apply “the data-based downside index” model on them. So far a short position was only held until day $t + 1$ if a trigger happened on day t , meaning that the effect of a trigger only lasted one day, then on day $t + 1$ the model was retrained involving the coming subsequent data point, and so forth. By restricting the holding period of the short position in a fixed one-day period, I unintentionally assumed that the effect of a signal, no matter its strength, is

always constant. Therefore, in these tests, I am going to lift this restriction and extend the modeling framework to investigate whether a pattern appears from changing the short-holding period. The tests would be performed on both previously determined significance levels (appointed based on classicist and rationalist views).

The signal strength is going to be measured in two alternative ways. In the first method, it will be represented as the number of keywords triggering at the same time. With the second method, I also intend to measure the signal strength directly, through the t-statistics of the separate univariate linear regressions. So I am also going to calculate the daily average of these t-statistics (which are above the confidence level), and multiply it with the number of words triggering, gaining another indicator for representing daily signal strength. Additionally, we obviously don't want to lengthen the holding period of the short position infinitely. Da et al. (2015) say that although the changes in their FEARS index correspond with the market movements, in the following few days this relationship tends to reverse. Tetlock (2007) made similar observations in his study. He highlights that the sentiment theory predicts returns on a short-horizon but in the long run markets tend to smooth out, thus, returns will be reversed. He experienced that the effect of negative sentiment is significant but only temporary, and the impact is usually fully reversed in one week. Following these observations, I will also apply a 7 days limit as the maximum length of the holding period of a short position. It is to be noted though, that the actual empirical holding period could be much longer if through an already held short position another significant trigger happens. Obviously, this should be considered as a separate event, therefore, if the strength of the new signal would indicate a longer shorting period than the one we are already holding at that time, then the shorting period would be updated, and lengthened, based on the new signal.

6 Model results

In this chapter, I will be showing how I have built my model gradually according to the research design I have devised earlier. I will be following the step-by-step procedure planned out there, in order to avoid any doubts about p-hacking, since my primary goal is to provide a well-documented and trustworthy methodology. To support the correct evaluation and comparison of the results, I will start with some technical specifications before showcasing the outcomes.

6.1 Technical specifications

To be able to assess the profitability of the model meaningfully I would be benchmarking my results to the “buy and hold strategy”. It is implemented by buying the index in the beginning and selling it at the end of the period. Since the models at a later stage

(beginning from the “data-based” approach) require an initial training period of data points, the evaluation period will start from 09/19/2004 until 28/08/2020, thus covering almost 16 years. The cumulative log yield for the benchmark “buy and hold strategy” is c.a. 102% in this period, which translates to an annual average of c.a. 6.4%. I have also calculated a lower limit floor based on the rounded U.S. inflation rate throughout the testing time frame. According to the reports of the U.S. government, the cumulative log rate is approximately 40% (U.S. Bureau of Labor Statistics, 2020), thus the annual average is 2.5%. This functions kind of like a break-even point: it shows the possible value growth of the money if it would not have been invested in the stock market. It is important to note that it should not be considered as an actual secondary benchmark rate, its purpose is to support the visualization of the distribution of the results (in some cases where needed). I will present most of my results in a heat map format for visibility purposes. If not stated otherwise, these values will be used as the mapping points of the coloring. All yields below 0% will be colored with red, between 0% and the annual average of the U.S. inflation rate (2.5% for the 16 years) I will apply the shades of yellow, and until the benchmark rate (6.4% for the 16 years) it will gradually take up the color green. If the benchmark is exceeded, it will be highlighted with a darker, more intense green color.

6.2 Results of the model development process

In this chapter, I will be presenting the results of the consecutive stages of the model development process. During the showcasing of the model outcomes, I will be following the same logic as described in the research design where the background and justification of each step are detailed.

6.2.1 Results of the “literature-based” approach

I have reconstructed the “literature-based” approach following the guidelines of Preis et al. (2013), then first I have performed the tests for the three matching keywords. This way, the only direct modification I have made compared to their framework was the utilization of daily data points instead of weekly ones. The results are summarized in Figure 6 below.

"literature-based" approach (until 22/02/2011)								
weeks	1	2	3	4	5	6	average yield (by keyword)	average yield of Preis et al. (2013)
debt	-0.4%	-8.4%	-5.5%	-3.0%	-1.0%	0.3%	-3.0%	32.3%
crisis	-18.1%	-8.4%	-8.6%	-5.4%	-3.8%	-6.6%	-8.5%	11.8%
unemployment	9.9%	9.1%	13.2%	11.6%	12.6%	11.4%	11.3%	18.7%
average yield (by week)	-2.9%	-2.5%	-0.3%	1.1%	2.6%	1.7%		

Figure 6: Results of the literature-based approach

Source: Own figure

On the left side of the figure, we can see how the three keywords performed with the moving average of 1, . . . , 6 weeks. I have applied the coloring method described in the technical specifications, with the only exception that the annual average log yield of the inflation rate and the buy-and-hold benchmark strategy was modified to 2.2% and 2.9% (respectively), after recalculating them for the period from 19/09/2004 until 22/02/2011. On the lower part of the figure, I have also calculated the average yield of the keywords per week. A trend seems to show up from these averages: the more past weeks we include in our moving average, the better performance we can achieve. However, I would advise against drawing serious conclusions from this, because not only are these averages based on only three keywords, the individual keyword level results are also very volatile.

The focus should be more on the comparison of the average yields of the reconstructed strategy and the reported results of Preis et al. (2013), which is presented on the right side of the figure. The differences are drastic. Not only did the modified strategy underperform compared to the results of Preis et al. (2013), in the case of the “debt” and “crisis” keywords we can only book huge losses. There can be two reasons why I got so different results compared to their article. Obviously one could state that this might be caused by the modified frequency of the data points, exchanging the weekly data points to daily ones. However, that should not matter this much, because eventually, the weekly data points are the results of aggregating the daily data, based on the documentation of Google Trends. So even though my strategy was run more frequently on a longer time frame, we should be expecting similar results (but at least similar trends) for the two frequencies. Consequently, the only thing that could lead to this big of a difference is that if Preis et al. (2013) did not apply a standardization method, or applied one which had look-ahead bias involved. However, to entirely safely make such a statement, we should need more testing, including fully replicating their modeling framework, using the same keyword pool with weekly search volume, and compare the effect of standardized and unstandardized datasets.

6.2.2 Comparison of the results of the “literature-based” and “dictionary-based” approaches

As a next step, I have rerun the same “literature-based” approach for the originally set time frame (until 08/28/2020), and simultaneously I did the same for the “dictionary-based” methods. I decided to present the results of the two approaches together in Figure 7 because obviously there are many similarities between the two. In the columns, the results of using averages of different lengths (1 to 6 weeks) can be seen, separately for both approaches. The keywords take place in the rows, categorized into what sentiment they have based on the Loughran-McDonald (2011) Financial Sentiment Dictionary. The average yields by weeks for all the keywords are shown in the lower part of the figure.

weeks	"literature-based" approach						"dictionary-based" approach					
	1	2	3	4	5	6	1	2	3	4	5	6
positive												
reforms	1.9%	-6.1%	-5.8%	-5.5%	-2.8%	-3.6%	-1.9%	6.1%	5.8%	5.5%	2.8%	3.6%
boost	6.4%	5.2%	9.7%	8.7%	8.6%	8.9%	-6.4%	-5.2%	-9.7%	-8.7%	-8.6%	-8.9%
consolidate	-8.5%	-7.3%	-10.5%	-10.2%	-9.5%	-9.2%	8.5%	7.3%	10.5%	10.2%	9.5%	9.2%
construction	-5.5%	-5.0%	-5.8%	-7.0%	-6.6%	-6.5%	5.5%	5.0%	5.8%	7.0%	6.6%	6.5%
outperform	-2.2%	-5.8%	-4.0%	-4.5%	-3.5%	-5.1%	2.2%	5.8%	4.0%	4.5%	3.5%	5.1%
savings	-8.4%	-6.1%	-3.6%	-3.0%	-2.0%	-2.2%	8.4%	6.1%	3.6%	3.0%	2.0%	2.2%
progress	-3.9%	-4.1%	-3.3%	-2.6%	-2.9%	-3.2%	3.9%	4.1%	3.3%	2.6%	2.9%	3.2%
booming	-11.0%	-5.1%	-7.0%	-6.9%	-9.2%	-10.9%	11.0%	5.1%	7.0%	6.9%	9.2%	10.9%
accrue	0.8%	0.8%	0.0%	0.3%	-0.2%	-1.0%	-0.8%	-0.8%	0.0%	-0.3%	0.2%	1.0%
surpass	5.5%	3.0%	0.7%	0.4%	0.8%	3.0%	-5.5%	-3.0%	-0.7%	-0.4%	-0.8%	-3.0%
negative												
debt	-1.3%	-3.4%	-2.3%	-0.8%	-0.8%	0.7%	-1.3%	-3.4%	-2.3%	-0.8%	-0.8%	0.7%
crisis	-8.3%	-3.8%	-2.2%	0.9%	0.3%	-2.8%	-8.3%	-3.8%	-2.2%	0.9%	0.3%	-2.8%
decline	-7.0%	-3.1%	-0.6%	-1.4%	-1.8%	-0.9%	-7.0%	-3.1%	-0.6%	-1.4%	-1.8%	-0.9%
recession	-2.7%	2.4%	0.4%	-1.8%	-3.4%	-3.1%	-2.7%	2.4%	0.4%	-1.8%	-3.4%	-3.1%
unemployment	8.9%	6.6%	7.9%	4.9%	3.6%	2.3%	8.9%	6.6%	7.9%	4.9%	3.6%	2.3%
bankrupt	-2.1%	3.1%	4.2%	3.6%	4.8%	5.0%	-2.1%	3.1%	4.2%	3.6%	4.8%	5.0%
deficit	1.6%	9.7%	7.0%	7.3%	7.8%	5.4%	1.6%	9.7%	7.0%	7.3%	7.8%	5.4%
collapse	-1.9%	-0.3%	-3.8%	-1.5%	3.9%	5.4%	-1.9%	-0.3%	-3.8%	-1.5%	3.9%	5.4%
market crash	-4.2%	6.6%	1.9%	0.4%	0.5%	1.3%	-4.2%	6.6%	1.9%	0.4%	0.5%	1.3%
downturn	-6.4%	-2.6%	-2.4%	-5.3%	-3.6%	-3.7%	-6.4%	-2.6%	-2.4%	-5.3%	-3.6%	-3.7%
neutral +												
banana	-4.1%	-6.2%	-1.9%	-3.0%	-1.7%	-3.9%	-4.1%	-6.2%	-1.9%	-3.0%	-1.7%	-3.9%
door	2.8%	0.8%	-0.6%	-2.9%	-1.2%	-3.2%	2.8%	0.8%	-0.6%	-2.9%	-1.2%	-3.2%
weather	-3.6%	-3.9%	1.4%	4.5%	5.4%	3.8%	-3.6%	-3.9%	1.4%	4.5%	5.4%	3.8%
chopstick	-3.3%	3.6%	1.9%	1.7%	1.4%	1.4%	-3.3%	3.6%	1.9%	1.7%	1.4%	1.4%
phone	6.5%	5.9%	8.1%	6.1%	7.0%	5.8%	6.5%	5.9%	8.1%	6.1%	7.0%	5.8%
chicago	1.5%	-2.3%	-5.5%	-2.3%	-2.0%	-1.8%	1.5%	-2.3%	-5.5%	-2.3%	-2.0%	-1.8%
christmas	-0.1%	3.6%	3.3%	4.1%	5.7%	3.1%	-0.1%	3.6%	3.3%	4.1%	5.7%	3.1%
swim	-1.8%	-2.4%	-1.6%	-0.6%	-1.6%	-2.3%	-1.8%	-2.4%	-1.6%	-0.6%	-1.6%	-2.3%
bycycle	-2.0%	2.9%	2.1%	1.3%	0.5%	0.7%	-2.0%	2.9%	2.1%	1.3%	0.5%	0.7%
programming	0.3%	4.2%	4.5%	7.1%	4.5%	3.4%	0.3%	4.2%	4.5%	7.1%	4.5%	3.4%
neutral -												
banana	4.1%	6.2%	1.9%	3.0%	1.7%	3.9%	4.1%	6.2%	1.9%	3.0%	1.7%	3.9%
door	-2.8%	-0.8%	0.6%	2.9%	1.2%	3.2%	-2.8%	-0.8%	0.6%	2.9%	1.2%	3.2%
weather	3.6%	3.9%	-1.4%	-4.5%	-5.4%	-3.8%	3.6%	3.9%	-1.4%	-4.5%	-5.4%	-3.8%
chopstick	3.3%	-3.6%	-1.9%	-1.7%	-1.4%	-1.4%	3.3%	-3.6%	-1.9%	-1.7%	-1.4%	-1.4%
phone	-6.5%	-5.9%	-8.1%	-6.1%	-7.0%	-5.8%	-6.5%	-5.9%	-8.1%	-6.1%	-7.0%	-5.8%
chicago	-1.5%	2.3%	5.5%	2.3%	2.0%	1.8%	-1.5%	2.3%	5.5%	2.3%	2.0%	1.8%
christmas	0.1%	-3.6%	-3.3%	-4.1%	-5.7%	-3.1%	0.1%	-3.6%	-3.3%	-4.1%	-5.7%	-3.1%
swim	1.8%	2.4%	1.6%	0.6%	1.6%	2.3%	1.8%	2.4%	1.6%	0.6%	1.6%	2.3%
bycycle	2.0%	-2.9%	-2.1%	-1.3%	-0.5%	-0.7%	2.0%	-2.9%	-2.1%	-1.3%	-0.5%	-0.7%
programming	-0.3%	-4.2%	-4.5%	-7.1%	-4.5%	-3.4%	-0.3%	-4.2%	-4.5%	-7.1%	-4.5%	-3.4%
average yield (by week)	-1.5%	-0.7%	-1.0%	-1.3%	-1.1%	-0.9%	0.0%	1.1%	1.0%	0.9%	1.0%	1.0%

Figure 7: Comparison of the results of the literature-based and dictionary-based approaches

Source: Own figure

Even after a quick look at the “literature-based approach”, we can once again see how hectic the different results are. We can see some high-performers on the individual level (such as “boost”, “unemployment” and “collapse”), however, the high-level picture seems

discouraging because with most of the keywords we not only did below the annual average buy-and-hold benchmark rate (6.4%), but we accumulated huge losses. If we take a look at the average performance of the keywords, we can't really identify any trends.

Moving onto the results of the “dictionary-based” approach, we can see a slightly increased average performance. Preis et al. (2013) tested all their keywords as if the attention directed to them had a strong negative relationship with the market, thus, as all their keywords had a strong and constant negative sentiment with the market. That is why it should not come as a surprise, that with the “dictionary-based” approach we got the same results as the “literature-based” results for the negative keywords and the neutral ones tested as negatives, and we got completely opposite results for the positive words and the neutral ones tested as positives.

We can clearly see an increasing performance altogether for the keywords with a positive sentiment, as it was expected originally. Also, altogether the neutral ones seem to be lagging behind the keywords which have an economic significance (either having positive or negative economic sentiment), however, the difference is not as striking as I have assumed beforehand. Even though the average yields by week increased above 0% for each week, we still cannot see any trends which could prove that either of the weeks would be doing significantly better than the other ones. This coincides with our earlier hypothesis, that probably the number of used weeks does not really contribute that much to the performance, and presumably, that is why Preis et al. (2013) did not emphasize this in their study. I have stated earlier, that no matter this result, to avoid p-hacking, we would be continuing our modeling with the 3 weeks' average. In this particular case, it seems that this decision does not have any serious consequences.

From the results, we can also see that there are certainly some keywords (e.g. “boost”), for which the “dictionary-based” approach severely underperformed, and ruined their rates compared to the “literature-based” method. This seems to be promising for our future “data-based” approaches because there we would like to precisely dodge these types of errors coming from the 'manual' predestination of sentiment.

6.2.3 Analysis of sentiment via the “data-based” approach

The “literature-based” approach strongly focused on capturing attention, and it simply presumed that the relationship of this attention would be negative and constant with the market. The “dictionary-based” approach introduced predefined sentiment into our models through the dictionary, although the strength of the sentiment was still not measured, it was assumed to be constant. With the “data-based” approach we are lifting all these limitations, and we let the data determine itself, and decide what the strength and the direction of a given keyword's sentiment is. For this purpose, we have used univariate linear regressions, following the guidelines of Da et al. (2015). Figure 8 summarizes my

findings on the strength and the direction of the sentiment of the keywords. For each keyword, I provided the average of the betas and the number of positive and negative betas contributing to this average on the 95% significance level on the left side, and for all cases (meaning no significance level was taken into account) on the right side.

dictionary-based sentiment		95% significance level			all cases		
		average of betas	number of betas		average of betas	number of betas	
			positive (+)	negative (-)		positive (+)	negative (-)
positive	<i>reforms</i>	0.0012	6	-	0.0002	3402	612
	<i>boost</i>	-	-	-	-0.0010	-	4014
	<i>consolidate</i>	0.0012	240	-	0.0007	4014	-
	<i>construction</i>	-	-	-	0.0003	2984	1030
	<i>outperform</i>	0.0065	286	-	0.0007	2687	1327
	<i>savings</i>	-	-	-	0.0003	2956	1058
	<i>progress</i>	-	-	-	0.0001	3057	957
	<i>booming</i>	0.0009	118	-	0.0003	2983	1031
	<i>accrue</i>	0.0046	111	-	0.0021	3958	56
	<i>surpass</i>	-	-	-	0.0001	1180	2834
negative	<i>debt</i>	0.0025	33	-	0.0010	3765	249
	<i>crisis</i>	-	-	-	0.0007	3951	63
	<i>decline</i>	0.0019	15	-	0.0005	4014	-
	<i>recession</i>	0.0011	628	-	0.0008	3951	63
	<i>unemployment</i>	-0.0011	-	245	-0.0001	2713	1301
	<i>bankrupt</i>	-	-	-	0.0001	3347	667
	<i>deficit</i>	-	-	-	0.0002	3033	981
	<i>collapse</i>	-	-	-	0.0002	3648	366
	<i>market crash</i>	-	-	-	0.0000	2278	1736
	<i>downturn</i>	0.0027	620	18	0.0005	3706	308
neutral	<i>banana</i>	-	-	-	-0.0015	93	3921
	<i>door</i>	-	-	-	0.0004	3451	563
	<i>weather</i>	-	-	-	-0.0005	1006	3008
	<i>chopstick</i>	0.0008	107	-	0.0005	3566	448
	<i>phone</i>	-	-	-	0.0005	2796	1218
	<i>chicago</i>	-	-	-	-0.0001	2945	1069
	<i>christmas</i>	-	-	-	-0.0005	215	3799
	<i>swim</i>	0.0022	26	-	0.0013	4014	-
	<i>bycycle</i>	0.0014	363	-	0.0005	3998	16
	<i>programming</i>	-	-	-	0.0009	3853	161

Figure 8: Analysis of sentiment via the data-based approach

Source: Own figure

Upon viewing the results, one should focus on the left column showcasing the test results on a 95% significance level. We can see that for many of the words we did not even get one single beta that was significant. What is even more stunning, is that the only keyword that resulted in a negative relationship with the market was “unemployment”. Keywords that should have a strong negative relationship towards the market (such as “debt”, “decline”, etc.) unambiguously acquired only positive betas. If we take the example of “decline”, we can see that even if we consider all the insignificant betas (the right column), we have

only positive values for a keyword that we originally assumed to be negative based on the dictionary. And we can see many more similar cases. Also, we should highlight that some of the neutral keywords seem to have an occasional stronger relationship with the market movements.

If the “dictionary-based” approach would be a valid and justified method, then here we should have gotten very different results. In that case, the words with an economic sentiment (the positive and negative ones) should have gotten average betas with matching signs, and the number of betas of that sign strongly surpassing the opposite. Furthermore, the neutral ones should have gotten almost no significant betas compared to the positive and negative terms. After our empirical test, we can clearly see that it is not the case and based on the results we can form two conclusions. First, the strength of the sentiment is clearly not constant over time. This is absolutely in line with my prior expectation. However, the second conclusion is rather surprising: the direction of the sentiment based on the dictionary is also misleading, and therefore, one should definitely apply some sort of data-driven method to confirm the strength and the direction of a keyword’s sentiment.

6.2.4 Results of the “data-based” approach

Following the descriptive analysis of the sentiment, we should move onto observing the performance of this strategy. In Figure 9, I have included all the keywords for which the “data-based” approach at least once declared economic significance (had significant betas), thus strong correlations with the market. For backtesting purposes (presented in Section 5.4.4), I also investigated whether the short positions or the long positions contributed more to the overall cumulative performance of a given keyword.

"data-based" approach				
dictionary-based sentiment		short & long	short only	long only
positive	reforms	0.05%	0.13%	-0.08%
	consolidate	0.07%	-0.07%	0.14%
	outperform	1.33%	1.98%	-0.65%
	booming	0.47%	-0.71%	1.18%
	accrue	-0.30%	-0.45%	0.16%
negative	debt	-0.24%	0.55%	-0.79%
	decline	0.23%	0.21%	0.02%
	recession	-1.59%	-0.57%	-1.02%
	unemployment	-0.36%	-0.27%	-0.09%
	downturn	1.86%	1.00%	0.86%
neutral	chopstick	2.00%	0.12%	1.88%
	swim	-0.23%	-0.10%	-0.13%
	bycicle	-2.28%	-1.44%	-0.84%
average		0.08%	0.03%	0.05%
standard deviation		1.22%	0.95%	0.69%

Figure 9: Results of the data-based approach

Source: Own figure

Figure 9 clearly shows that the results of this strategy are not on the same scale as the previously tested strategies, the performance of each keyword individually, and thus the average as well tends to oscillate very close to 0% (that is the reason why I was forced to present the yields with two decimals). This is presumably due to the fact that we are spending most of our time in a null position. More precisely, by calculating how many times we have a significant beta for these words based on Figure 9, we can see that only c.a. 5.6% (226/4014) of all days did we have either a long or a short position on average. For the rest of the time, we are incapable of earning yields in this simplified modeling framework. This is according to our prior hypothesis, and that is exactly the reason why we moved on eliminating the null position from our model with the “data-based downside” approach later.

Additionally, we can see the distribution of the performance between the short and long positions on the right side of the figure. An evident trend does not seem to be emerging from these results, the distribution is rather hectic. The average and the standard deviation of the short only and long only results also approximately match with each other. Therefore we can't really confirm nor deny the observations of Tetlock (2007) and Da et al. (2015) whether the negative sentiment would be a stronger signal than the positive one.

6.2.5 Results of the “data-based downside” approach

Next, we have introduced the “data-based downside” approach that had the primary goal to eliminate the null position from our model. With this approach, we are only focusing on the negative sentiment resulting in short positions, and if that does not trigger, we remain in a long position as a default. We have already seen that for most of our keywords the regression model does not signal any economic significance at all during the whole testing period, therefore, I did not include them in Figure 10.

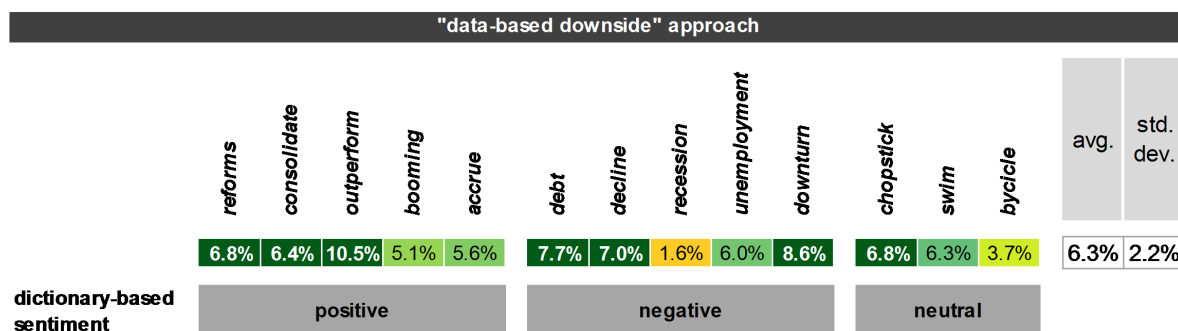


Figure 10: Results of the data-based downside approach

Source: Own figure

The average of the yields included in the table is approximately in line with the annual average buy-and-hold rate (6.4%), but we can clearly see some individual high-performers (such as the word “outperform” with a 10.5% annual average yield). However, I would advise against assessing these results solely on an individual level, we should rather concentrate on the high-level picture. The progress compared to the simple “data-based” approach is enormous. Based on the modified technical specifications (having the null position eliminated from the model), these results are now also comparable with the “literature-based” and “data-based” approaches. From a higher-level perspective, we can definitely see a significant increase in performance compared to both of these earlier approaches. All of our cumulative yields are positive for each keyword individually, including those as well which are not included in this figure: since there were no short positions triggered at all, we only kept the long position for the entire time, thus, we basically followed the buy-and-hold strategy, and earned yields accordingly.

6.2.6 Analysis of the economic significance of the keywords via the “data-based downside index” approach

By implementing the “data-based downside index” approach I have arrived at the final step of developing my modeling framework. So far all the methods I have investigated and evaluated remained on a keyword level, this is the first approach where a collective decision is formed based on the index of individual keywords. The biggest question when

constructing an index is what to include in it. We have already investigated the dictionary-based sentiment of the keywords, but the results looked discouraging. However, as I have mentioned in Section 5.4.5, there is an additional aspect of the Loughran-McDonald (2011) dictionary, which we haven't looked at yet but might be able to help to decide what keywords we should incorporate into an index: the dictionary also tells us which keywords are economically significant.

My assumption is that the ones that are presented in the dictionary (the positive and negative keywords) will outperform any of the other combinations of words. Therefore I am calling this the final model in Figure 11. The alternative approaches are listed on the right side, representing all the different groupings of my keywords. As I have already declared earlier, I am going to test my models following both the classicist and rationalist approaches regarding the determination of the confidence level.

"data-based downside index" approach									
		final model	models with alternative groupings					Legend: positive (+) negative (-) neutral (o)	
grouping via dictionary-based sentiment		+-	+	-	o	+o	-o		+ - o
number of keywords		20	10	10	10	20	20		30
Classicist	Sig. level	99.75%	99.5%	99.5%	99.5%	99.75%	99.75%		99.83%
	Returns	9.6%	6.6%	9.5%	6.2%	6.6%	9.6%		9.0%
Rationalist	Sig. level	98%							
	Returns	11.1%	8.8%	8.4%	4.5%	6.9%	6.4%	9.3%	

Legend:

positive (+)

negative (-)

neutral (○)

Figure 11: Analysis of the economic significance of the keywords

Source: Own figure

After looking at all the model results from a high-level perspective, we can conclude that all the models did fairly well. There is no doubt that the model based on only the neutral keywords, thus the keywords that do not have any economic significance based on the dictionary, lag the farthest behind all the others, for both the classicist and the rationalist approaches. All the other alternative models, where we had at least the positive or the negative keywords included (even when they form separate groups on their own), performed really sufficiently. In the majority of the cases, we can see that they outperformed the benchmark rate of the buy-and-hold strategy or they got really close to it. However, for both p-value adjustment approaches the final model outperformed all the alternative approaches, including the benchmark rate as well.

Based on these results our original hypothesis seems to stand: although the sentiment of the dictionary did not prove to be reliable, the represented economic significance has a significant effect when forming our index. That also indirectly confirms our earlier assumptions, whether a data-driven method should be used to conclude the strength and

the direction of sentiment. Before I take a deep dive into analyzing the final model further, I will present the test results regarding the analysis of the strength of this collective signal.

6.2.7 Analysis of the collective signal strength via the “data-based downside index” approach

The signal strength is represented in two alternative ways. First, the signal strength is measured by the number of words that triggered for a given day. In the second approach, I have also taken into account a triggering keywords’ strength of the relationship with the market, thus, I have multiplied the number of triggering keywords with the absolute value of the average of their respective t-statistics, which is ultimately equivalent with the aggregated t-statistics of the triggering keywords. I have tested these approaches against different days of holding a short position, following the arguments of Da et al. (2015) and Tetlock (2007) presented in the research design. I have summarized the results into tables formatted into heat maps to give a clearer understanding regarding any trace of correlation between the signal strength and the length of the short-holding period. To avoid mixing the previously used benchmarking logic with this one, I used a different coloring template: the warmer the color gets, the higher the performance is for the given short-holding period and signal strength pair, thus, orange means a relative high-performer, while blue means a relative low-performer. In the Appendix, the number of transactions for each of the performance figures is also included (please refer to Figure 16 to the Appendix). I will first present the results of the first then the second way.

		short-holding period (days)						
number of keywords	98% sig. level	1	2	3	4	5	6	7
	1	11.9%	11.0%	9.9%	10.5%	9.5%	8.9%	9.4%
	2	5.7%	6.0%	6.9%	6.3%	6.2%	5.7%	6.1%
	3	6.7%	6.4%	6.5%	6.0%	6.3%	6.1%	5.9%

		short-holding period (days)						
number of keywords	99.75% sig. level	1	2	3	4	5	6	7
	1	9.6%	9.3%	8.0%	9.5%	9.7%	9.8%	9.5%
	2	6.4%	6.4%	6.4%	6.4%	6.4%	6.4%	6.4%
	3	6.4%	6.4%	6.4%	6.4%	6.4%	6.4%	6.4%

Figure 12: Analysis of the collective signal strength - first approach

Source: Own figure

The results of the first way (shown in Figure 12) of representing signal strength are not so apparent. Unfortunately, due to the height of the significance level for the classicist approach, we could not observe any keywords triggering at the same time. On the other hand, the rationalist approach gave more interpretable results, but it does not seem to show any comprehensive trends. Although the upper left part of the table seems to show that the lower the signal strength the lower the short-holding period should be, the continuation of this pattern can’t be observed for the higher signal strength and higher short-holding period.

Both tables (both the classicist and rationalist one) indicate that choosing the number of

triggering keywords was not an ideal way of assessing the signal strength. For the classicist approach, we don't have more than one keyword triggering at the same time, and even for the rationalist approach, our maximum number of simultaneously triggering keywords is three. Due to this, the tables are vertically not wide enough, the measured performances are not scattered enough, and the concentration of the values is too high. Therefore, I would advise not to draw serious conclusions from this representation.

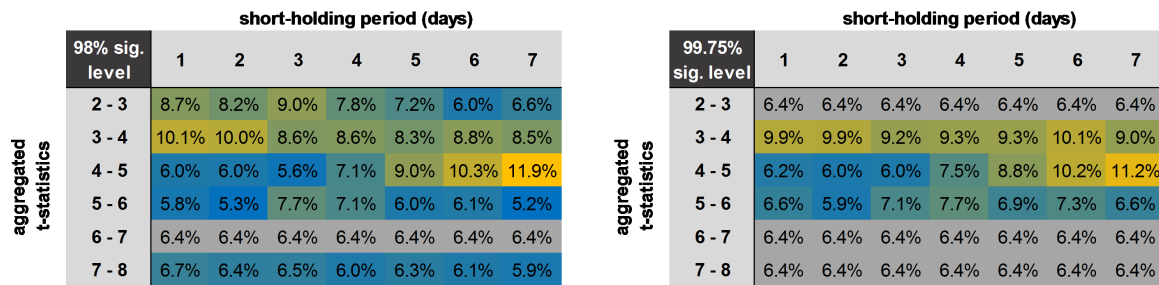


Figure 13: Analysis of the collective signal strength - second approach

Source: Own figure

The second way (represented in Figure 13) seems to fix the issues experienced by the first one, here the tables have similar dimensions vertically and horizontally, and even for the classicist approach the results got more scattered. By interpreting the heat maps, we can point out that both tables seems to show a similar trend: the higher the signal strength grows, the bigger the short-holding period should be. Based on the lower parts of both tables we can't safely state anything, because the number of transactions are too low here, such strong signals are measured quite rarely.

The above results support our assumption that signal strength is positively correlated with the short-holding period. Although the outcome of the analysis seems to be really promising, it is important to note that this is only a first step of an explanatory research towards actually able to confirm such a grand hypothesis. Therefore, hopefully in the future, these results can become a well-grounded basis for starting a comprehensive research to explore this direction of Google search volume prediction.

6.3 Evaluation of the performance of the final model

As a final step, I would like to further reflect on the performance of the final model. After iterating over several modeling decisions, beginning from the model of Preis et al. (2013) (named as the "literature-based" approach) and building in the findings of Da et al. (2015) among several other additions, we have eventually reached the final stage of our model. The final model combines both the attention-measuring and the relationship-measuring factors, creating a comprehensive framework to determine investor sentiment by using Google

search volume. Not only did we make improvements towards gaining a more dynamic and prudent technical framework, but as the results of the different stages indicated, we could also gradually raise the performance of the model. Our final model outperformed the average annual 6.4% yield of the benchmark buy-and-hold portfolio on both used significance levels, achieving 11.1% annual yield on 98% significance level and 9.6% annual yield on 99.75% significance level. Just to give a sense of the magnitude of the difference, this means that over the almost 16 years long test period the c.a. 102% cumulative yield of the benchmark portfolio was beaten by the model scoring approximately 152% and 177% cumulative yields on the 98% and 99.75% significance levels respectively.

To achieve these yields we had to place 456 and 100 transactions altogether, working with 98% and 99.75% significance levels respectively. These numbers include taking the opening position at the start of the testing phase, and closing the final position when we are theoretically realizing the earnings on the final day. Please also note that every time we are changing positions two transactions are placed, one for closing the current positions, and simultaneously, one for opening an opposite position. Additionally, it is important to highlight here, that no transaction costs nor the bid-ask spread were taken into account during our modeling work. So although the number of transactions seems to be reasonably moderate, and the achieved lead in comparison with the benchmark portfolio is tremendous, in an actual trading scenario these factors would reduce our net earnings.

On average, when we decided to take up a short position, we have acted on 1.16 and 1.00 words, on the 98% and 99.75% confidence levels respectively. For the 98% level maximum of three words triggered simultaneously, while on the 99.75% level there were no occasions where more than one word signaled on the same day. This can be seen also as a counterpoint to the classicist p-value determination: since our model performed slightly better on the lower significance level, by raising the significance level too harshly, in order to minimize the chance of Type I errors, we have missed some valuable opportunities.

I also wanted to take a closer look, at which periods did the final model predict a short position, thus, where exactly the additionally earned yield comes from compared to the buy-and-hold strategy. Therefore, I visualized the time frame and the short positions in Figure 14. I have also added the price of the DJI during this period for better visualization purposes. From the signal strength analysis, the number of keywords multiplied by the average t-statistic seemed to be an adequate indicator of signal strength, therefore I decided to represent the short positions accordingly (this is only an extra feature, so no conclusions should and will be drawn from this). Please note that whenever a short position is triggered on the 99.75% significance level, it was also obviously triggered on the 98% significance level, thus technically, the red lines should also be counted as orange ones.

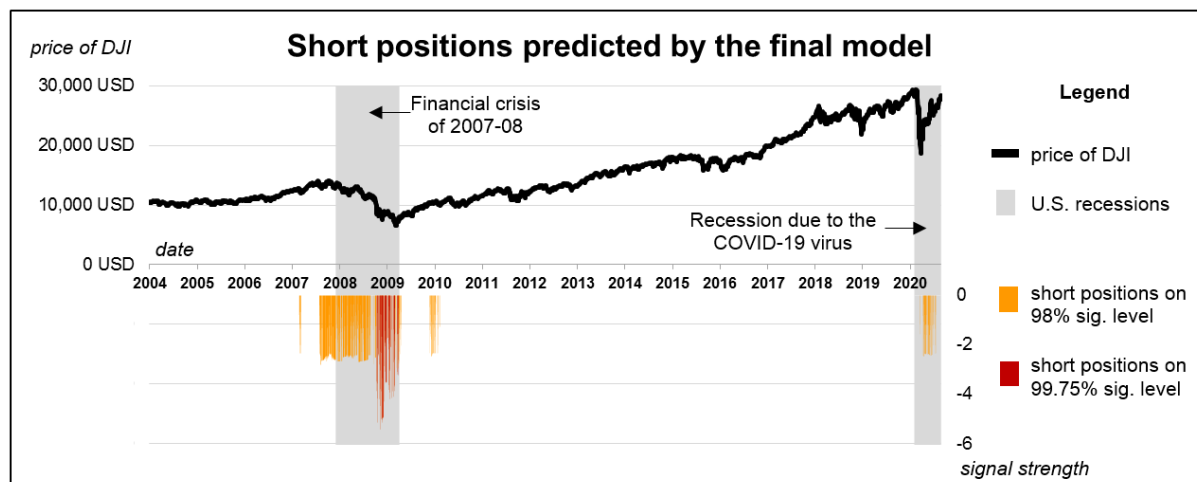


Figure 14: Short positions predicted by the final model

Source: Own figure

Figure 14 gives a clear illustration of why our model performed in such an outstanding way: most of our predicted short positions overlap with the most severe U.S. recessions, with the financial crisis of 2007-08 and the recent market falls induced by the COVID-19 virus. On the 99.75% confidence level, our model triggered short positions for the second half of the financial crisis of 2008-09, and we have realized huge profits during this period. Although the model could not forecast the most adequate entry point for shorting the whole crisis, it is extremely interesting how punctually it predicted when the market trends changed, ceasing all shorting activities in time. On the 98% significance level, our model did even better. In addition to the triggers we could observe previously on the 99.75% confidence level, we spent almost the entire financial crisis of 2008-09 in short positions, only occasionally breaking the streak of those. What is especially fascinating, that it had already indicated and taken up several short positions months before the actual crisis hit. Moreover, not only did it manage to forecast the financial crisis so well, but it also pinpointed the recession due to the COVID-19 virus very accurately, earning even more profits from the sudden and large falls of this period. We can see that other than these two occasions, the triggers are quite rare. All in all, our model is performing really well, and it seems to be due to predicting the major recessions precisely and in time.

7 Robustness testing

We could see from the previous chapter that the final model achieved very promising results, and it could predict the major U.S. (and global) recession periods quite precisely. However, to assure the quality and the stability of the methodology and the results, moreover, to assess the true potential of this modeling framework, the robustness of the outcomes should be verified.

7.1 Proposing the robustness test

As the main edge of our final model appears to be its remarkable ability to forecast and perform well during times of recessions, I am going to propose a robustness test that investigates exactly this characteristic. There are several ways to perform this test, all involving replacing the input search volumes or the to-be-predicted log yields with different variables. Ideally, for the sake of the robustness test it would be a good solution, to move away from the U.S., and repeat the whole exercise with the search volume and a representative index of another large country (such as China, Russia, the United Kingdom, Germany, etc.), or even with global search volume. However, there are two issues with this approach. The bigger problem is that this would mean several additional search volume downloads, which proved to be the most serious bottleneck of the whole research (it took 30 servers and almost 5 days to download the data for this exercise only). Another issue is that the other large countries, e.g. China, don't prefer using Google as their primary search tool. While Google is dominating the search engine market of the U.S. throughout this period, this sweeping majority might be not true for other countries as well, therefore Google search volume might not represent the belief and sentiment of the small investors of that given country.

Therefore, I am proposing a robustness test, which still uses the U.S. search volume, but we are replacing the Dow Jones Industrial Average index. We originally chose the DJI index to represent the U.S. stock market, and thus the U.S. economy, for several reasons. On one hand, it is a commonly followed index and is considered as one of the adequate indices to represent the overall U.S. stock market by the public as well. On the other hand, this index was also used in the paper of Preis et al. (2013). Nonetheless, these arguments can be challenged, therefore, I propose to test the robustness of my model on the collection of indices which are representing different sectors of the U.S. economy. According to my assumption, if my model is robust enough, it will excel in the case of most sectoral indices as well, because in the cases of abnormal events, when significant euphoria or panic is present (especially in the case of a crisis or recession), markets tend to move together as a whole. As a reason, we are likely to see the model perform on the different sectoral indices very similarly to its performance on the DJI.

For the above purpose, I am going to be using the most prominent funds that represent one of each of the 11 S&P sectors of the U.S. economy. It seems plausible to use the corresponding S&P funds to describe these sectors but in the case of two sectors, these funds are inadequate due to a much shorter time frame. The XLRE fund of the Real Estate sector was launched after mid-2015, and the XLC fund of the Communication Services sector began to trade only in mid-2018. To replace these, I brought on two corresponding Vanguard funds. Both of these were launched on 09/29/2004 which is still slightly later than the beginning of 2004 but since the whole testing period is almost 16 years long, this

small discrepancy can be considered irrelevant. The funds, used to represent each sector, are listed in Table 3 below. For the descriptive statistics of the datasets please refer to the Appendix to Table 5.

Sector	Representative fund	Ticker
Real estate	Vanguard Real Estate ETF	VNQ
Technology	Technology Select Sector SPDR Fund	XLK
Materials	Materials Select Sector SPDR Fund	XLB
Health care	Health Care Select Sector SPDR Fund	XLV
Industrial	Industrial Select Sector SPDR Fund	XLI
Consumer staples	Consumer Staples Select Sector SPDR Fund	XLP
Energy	Energy Select Sector SPDR Fund	XLE
Consumer discretionary	Consumer Discretionary Select Sector SPDR Fund	XLY
Communication services	Vanguard Communication Services ETF	VOX
Utilities	Utilities Select Sector SPDR Fund	XLU
Financial	Financial Select Sector SPDR Fund	XLF

Table 3: U.S. sectors and their representative funds

Source: Own table

7.2 The results of the robustness tests

I have applied the final model to the daily log yields calculated from the above mentioned sectoral indices. The model was run on both 98% and 99.75% significance levels, thus, it was tested by following both the rationalist and classicist p-value determination principles. I have summarized the results in Figure 15. In the case of each index, I have highlighted whether the buy-and-hold or our applied trading strategy (the final model) performed better, on both confidence intervals separately. The number of transactions needed to realize the active strategy is also shown.

	buy-and-hold	final model on 98% significance level		buy-and-hold	final model on 99.75% significance level	
ticker	average annual yield	average annual yield	number of transactions	average annual yield	average annual yield	number of transactions
VNQ	6.5%	6.8%	676	6.5%	8.7%	182
XLK	12.9%	13.1%	462	12.9%	12.8%	16
XLB	7.6%	4.2%	866	7.6%	7.0%	196
XLV	9.8%	9.0%	954	9.8%	10.7%	42
XLI	8.2%	8.8%	964	8.2%	10.4%	58
XLP	9.3%	6.2%	436	9.3%	9.7%	4
XLE	2.8%	8.6%	1206	2.8%	4.6%	222
XLY	11.1%	11.8%	666	11.1%	12.3%	302
VOX	7.2%	4.3%	462	7.2%	8.5%	4
XLU	9.0%	8.1%	8	9.0%	8.5%	4
XLF	3.9%	-2.1%	826	3.9%	4.9%	54
average	8.0%	7.2%	684	8.0%	8.9%	99
std. dev.	2.9%	4.1%	331	2.9%	2.7%	107

Figure 15: Results of the robustness tests

Source: Own figure

On the 98% significance level, the individual performances are rather scattered between the final model and the buy-and-hold strategy. However, we should be more focused on the average of the average annual log yields of the models, because the average performance of the most important sectoral indices represents the market as a whole. By comparing the average of the two, we can see that the buy-and-hold model altogether achieved bigger yields. It also did this in a more consistent way, since the standard deviation of the results is less than the one calculated from the individual outcomes of the active strategy. Also, the number of transactions seems to be rather high here. On the other hand, our observations made on the final model computed with a 99.75% confidence level are quite the opposite. The individual and average performance both show the final model as the clear winner, even the standard deviation shows a less volatile average output. Unsurprisingly, the number of transactions also decreased significantly.

All in all, the robustness test on the 99.75% significance level (p-value determined by the classicist approach) outperformed the benchmark, while the same test failed on the 98% significance level (p-value determined by the rationalist approach). However, the average difference in both cases is not that robust as we could see by the DJI, therefore, the interpretation of these results is not that straightforward. The rationalist approach achieved a tremendous lead when we were testing on the DJI, and it accomplished similar outcomes in the case of a few individual sectoral indices, but it slightly underperformed on average in the robustness test. The classicist approach also outperformed by scoring almost 1.5-fold more than the benchmark when working with the returns of the DJI, and it also concluded much better individual and average results than the buy-and-hold portfolio

in the robustness tests. Based on only these tests, I would advise against unequivocally and permanently devaluing the final model paired with the rationalist approach, but we can't neglect the results of the robustness test: the final model on a higher confidence level (via the classicist approach) seems to be a safer and more consistent strategy.

Finally, I would like to highlight the importance of robustness testing. By Preis et al. (2013) we could see much higher results compared to my final model, they achieved 326% cumulative yield over less than 8 years, while Zhong and Raghieb (2019) achieved almost 500% in 10 years (who used a very similar methodology to the former authors). These translate to approximately 45% and 50% average annual yields respectively. Comparing these to my 11.1% and 9.6% yields (on 98% and 99.75% confidence levels respectively), which still outperformed the benchmark buy-and-hold portfolio significantly, the difference is clearly striking. However, I find it quite surprising that in the paper of Preis et al. (2013), and in other similar researches such as the one of Zhong and Raghieb (2019), no robustness tests were performed at all. Additionally, we have uncovered many biases that could have affected those results, which increase the need for such validation even more. Therefore, when a model is presented without the proper justification of its robustness and trustworthiness, I would advise being careful when interpreting its results.

8 Potential model development areas

The following section is about the main points which should be thoroughly considered if we want to leave the theoretical scenario behind and utilize the model in real-time day-to-day trading. I also intend to point out some further areas of research, which could be beneficial to gain a deeper understanding of what more fundamental characteristics Google search volume prediction has. In all of the cases, I intend to give some recommendations on practical solution options based on my experience. I have summarized only the main potential model development areas here, I have already named a few further research directions in Section 5 and 6 when planning and evaluating my modeling framework.

8.1 Accounting for transaction costs and the bid-ask spread

Through my research, I did not put much emphasis on the question of transaction costs. A few years ago trading without transaction costs was only an assumption of the theoretical efficient-market hypothesis but these days this is very real. With Robinhood³ as a flagship, more and more brokerages and fintech startups offer free intermediary services for trading securities, hence I did not account for transaction fees in my model. However, if such a

3. Robinhood Markets Inc. offers a mobile app and website that gives people the ability to invest in stocks, ETFs, options, and several cryptocurrencies without charging any commissions or transaction fees. Due to the new fintech startup established players were forced to rewrite their business models by abolishing commissions as well (Egan, 2019).

need occurs, the model could be easily tweaked to accommodate this need, by involving and minimizing the number of transactions as an input parameter in the model. That is the reason why I tried to always attach and refer to the number of transactions when interpreting a model output. By keeping the number of transactions on a minimal level, the transaction costs would be obviously minimal as well.

The other aspect is the accounting for the bid-ask spread. As I highlighted before, the log yields were calculated from the adjusted closing price of the DJI. However, since our strategies are based on buying and selling the index multiple times, it would have been more precise to take the bid-ask spreads into consideration. For theoretical research, considering only the adjusted closing price is a completely acceptable constraint, because spreads vary between brokerages and financial service providers. Nonetheless, this limitation should be lifted before putting the model into practice. By doing so, spreads would act like transaction costs when we were to change our positions, therefore, the same approach, the minimization of the number of transactions, could be applied here as well.

8.2 Grammar- and trend-focused analysis of the keywords

I have presented an objective method of keyword selection by choosing keywords with positive and negative sentiment based on the Loughran-McDonald (2011) dictionary and a control group of neutral keywords. However, we could see that for the majority of the cases the dictionary-based sentiment was overruled by the data-driven sentiment-capturing approach, thus, the search volumes of the majority of the keywords were not having the expected relationship, both strength- and direction-wise, with the market movements. Although we have made this observation, we have no ground and evidence to explain the reason for it.

My proposal would be to do some further analysis on the keywords themselves. In my view, first, a comprehensive grammatical examination could take place to check whether the word-class or syntax makes any difference when their search volume is used in a prediction model. This way we could gain a picture of whether nouns, verbs, adjectives, etc. have a significantly bigger predictive power than all the other word-classes. Additionally, for example, in the case of verbs, the separate tenses could be also tested, for adjectives the comparative and superlative forms could be tried as well, while for nouns the difference between singular or plural forms could be investigated. Secondly, a trend analysis could be performed regarding the keywords. It could be examined in what years or months a keyword was popular and widely used, thus, we could account for the survivorship bias of the keyword selection. Additionally, by filtering out keywords that were not favored, or not even known in certain time periods (like “Bitcoin” before 2009), we could avoid introducing look-ahead bias into our model via the keyword pool.

8.3 Testing and documenting the implications of the standard deviation

One of the most yet unknown and untapped areas of Google search volume prediction is the sampling extraction method of Google Trends. This feature causes a lot of trouble since different realizations of the same dataset are not entirely the same, there is some standard deviation between these time series. In Section 4.1.4, I have already presented a few articles which confirm that due to the high cross-correlation (around 99%) between the different realizations this phenomenon is negligible. However, as I have also pointed out, the background and the documentation supporting these important statements are mostly rather laconic, therefore, there is still room for improvement in this area.

The first question is how different the standard deviation is for different keywords, but answering this question requires extremely complex research. On one hand, it should be checked whether the popularity of a keyword, which is then represented in its absolute search volume, contributes significantly to the magnitude of the standard deviation. Indirectly, this way the representativeness of Google Trends' sampling method would be confirmed, because if the sampling is representative enough, then the magnitude of the attention, thus the absolute search data at hand, should not matter that much. On the other hand, one also has to deal with the issue of how far back should the standard deviation be tested. It is not straightforward to decide what time frame we want to use. The whole historical time span means obviously more data to test on but that also implies more data requests. But if we only decide to test on a smaller time frame then we could never be sure about the overall robustness of the results. Additionally, the difference between different periodicities should be also accounted for, thus, the difference between daily, weekly and monthly frequencies.

The case of the overlapping time frame standardization method (presented in Section 5.2.4) is also interconnected with the representativeness of the sampling. If the standard deviation between different realizations is negligible, then the length of the overlapping part, used to calculate the average ratio, could be also minimized. However, if it turns out that the standard deviation has a decisive impact then we should be using the lengthiest overlap we are able to. In order to be prudent, I used the latter approach throughout my research, but it would have spared a lot of extra data requests if the robustness of the sampling method would have been unambiguously confirmed beforehand. It is important to note that to have a conclusive and robust test of the representativeness of the sampling method, thus the impact of the standard deviation, an enormous amount of data would be needed (historical datasets of as many keywords as possible), so we should not be surprised that this area remained yet mostly untapped.

8.4 Widening the scope of the assets

Our sentiment-capturing model was developed on the DJI, and its robustness was tested against several sectoral indices. The pool of our downloaded keywords limits in which areas the search volume at our disposal could be utilized. However, I strongly believe there is still some untapped potential of this particular model (with this particular pool of keywords) which could be harnessed. First of all, gold has been historically a safe haven for retail investors. Its price is proved to be in a strong negative correlation with stock market prices, revealing the true state and the actual health of the U.S. economy. Since our model showed a promising ability to forecast recessions, it could be tested against the gold price. Secondly, by taking a step back, we could investigate how well our model captures attention, and assess it against volatility indicators. I would propose using volatility measures that are representing the whole market, such as the Chicago Board Options Exchange Volatility Index (VIX) derived from the price inputs of the S&P 500.

If we could afford to download more search volumes, then a wide range of options and research directions open up. We can continue to work with the proposed sentiment-capturing model to make a similar attempt on forecasting the stock market movements of other countries or regions based on their own search volume. Or we could change the keyword selection approach, and conduct a bottom-up, industry-specific analysis by collecting keywords inherent to a given industry and running their search volume against a given sectoral index. It is to be noted that in each of these examples the already carefully developed and documented methodological framework could be applied, only the input parameters (the Google search volumes and the assets in scope) should be modified.

9 Conclusions

As a research topic, I wanted to reflect on the most acknowledged papers examining how Google Trends data can help quantify investor behavior in financial markets, especially focusing on the studies of Preis et al. (2013) and Da et al. (2015). After thorough literature research, I was quite surprised to see that the most popular papers, which achieved extraordinary results and simultaneously became modeling templates for the later researchers, have obvious methodological shortcomings. Among these deficiencies, we can mention the absence of validation, the noticeable statistical biases, the lack of dynamic model characteristics, and presumably even the need for standardizing the search volume datasets passed unmarked in the case of a few articles. I found this very alarming since one can get fascinated by looking at only the huge results, build false hopes and unachievable expectations, then assess a whole field of study (meaning all other papers) wrongly based on those.

When conducting this research, first I mapped all these possible biases and limitations, and

eventually, I have arrived at my final model by successfully introducing many new solutions and extending many model dimensions, following multiple other best practice examples. By doing so, the methodology became not only fully documented and transparent, but more refined and prudent too, thus, making the results more reasonable. Throughout the whole research process, another of my primary goals and considerations was that this model could be implemented into actual real-life day-to-day trading. I successfully managed to keep this intention of mine in parallel with adhering to other technical and statistical requirements.

The results of my final model, following the “data-based downside index” approach, look highly promising. On both tested significance levels (98% and 99.75%) it managed to considerably outperform the benchmark: over the 16 years from 2004 until mid-2020 an average annual yield of 11.1% and 9.6% was achieved (respectively), while the annual average yield of the buy-and-hold portfolio was c.a. 6.4%. I also found evidence that the final model forecast the U.S. recession periods quite precisely, earning the majority of the cumulative yields in these times, taking up short positions based on the search volume. This outcome seems to be proving the usability and effectiveness of using Google search volume to predict stock market movements, thus, it justifies my wider research question. Additionally, I believe I have reached several other milestones that contribute to the whole research area. One of the most substantial contributions of my work was mapping and introducing the standardization methods of the search volume datasets. By following the described steps, one could link together independently downloaded realizations, and make them comparable and standardized across time, without unintentionally including any look-ahead bias. To my knowledge, my paper was the also first one that defined and distinguished between investor attention and investor sentiment, including the different approaches with which these could be captured and utilized. Furthermore, the step-by-step research design gives a comprehensive guideline for future researchers, enumerating the key points that are fundamental for developing a search volume prediction model.

In this paper a justified and well-founded modeling framework was designed and documented, that challenged the status quo of this field of study. I sincerely hope that by inspiring others to do similarly, it can contribute to the further development of the highly promising search volume based modeling approach. All in all, the results and the robustness tests indicate that the model has good potentials for predicting stock market movements: even if decisions are not solely made on the basis of it, the model could have a place amongst the investment decision supporting tools.

References

- Askitas, N., and Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting. viewed 21 January 2020,
<<https://www.econstor.eu/bitstream/10419/35733/1/605353115.pdf>>
- Audrino, F., Sigrist, F., and Ballinari, D. (2020). The impact of sentiment and attention measures on stock market volatility. *International Journal of Forecasting*, **36**(2), 334-357. viewed 12 February 2020,
<<https://www.sciencedirect.com/science/article/pii/S0169207019301645>>
- Baker, S., and Fradkin, A. (2011). What drives job search? Evidence from Google search data. *Discussion Papers*, **10**-020. viewed 22 January 2020,
<<http://www.siepr.stanford.edu/RePEc/sip/10-020.pdf>>
- Bangwayo-Skeete, P. F., and Skeete, R. W. (2015). Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. *Tourism Management*, **46**, 454-464. viewed 3 April 2020,
<<https://www.sciencedirect.com/science/article/abs/pii/S0261517714001460>>
- Bewerunge, F. (2018). Google Trends: How to acquire daily data for broad time frames. viewed 9 December 2019, <<https://medium.com/@bewerunge.franz/google-trends-how-to-acquire-daily-data-for-broad-time-frames-b6c6dfe200e6>>
- Black, F. (1986), Noise, *The journal of finance*, **41** (3), 528-543. viewed 27 December 2019, <<https://onlinelibrary.wiley.com/doi/full/10.1111/j.1540-6261.1986.tb04513.x>>
- Böhme, M. H., Gröger, A., and Stöhr, T. (2020). Searching for a better life: Predicting international migration with online search keywords. *Journal of Development Economics*, **142**, 102347. viewed 24 September 2020,
<<https://www.sciencedirect.com/science/article/pii/S0304387819304900>>
- Bollen, J., and Mao, H. (2011). Twitter mood as a stock market predictor, *Computer*, **44**(10), 91-94. viewed 27 December 2019,
<<https://www.sciencedirect.com/science/article/pii/S187775031100007X>>
- Brown, G. W. (1999). Volatility, sentiment, and noise traders. *Financial Analysts Journal*, **55** (2), 82-90. viewed 27 December 2019,
<<https://www.jstor.org/stable/pdf/4480157.pdf>>
- Brown, S. J., Goetzmann, W., Ibbotson, R. G., and Ross, S. A. (1992). Survivorship bias in performance studies. *The Review of Financial Studies*, **5**(4), 553-580. viewed 30 October

2020,

<<https://academic.oup.com/rfs/article/5/4/553/1590264>>

Brownstein, J. S., Freifeld, C. C., and Madoff, L. C. (2009). Digital disease detection—harnessing the Web for public health surveillance. *The New England journal of medicine*, **360**(21), 2153. viewed 4 April 2020,

<<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2917042/>>

Challet, D., and Ayed, A. B. H. (2013). Predicting financial markets with Google Trends and not so random keywords. viewed 7 March 2020,

<<https://arxiv.org/pdf/1307.4643.pdf>>

Choi, H., and Varian, H. (2012). Predicting the present with Google Trends, *Economic Record*, **88**, 2-9. viewed 28 December 2019, Cooper, C. P., Mallon, K. P., Leadbetter, S., Pollack, L. A., and Peipins, L. A. (2005). Cancer Internet search activity on a major search engine, United States 2001-2003. *Journal of medical Internet research*, **7**(3), e36. viewed 23 January 2020,

<<https://www.jmir.org/2005/3/e36/>>

Da, Z., Engelberg, J., and Gao, P. (2011). In search of attention. *The Journal of Finance*, **66**(5), 1461-1499. viewed 3 February 2020,

<<https://onlinelibrary.wiley.com/doi/full/10.1111/j.1540-6261.2011.01679.x>>

Da, Z., Engelberg, J., and Gao, P. (2015). The sum of all FEARS investor sentiment and asset prices. *The Review of Financial Studies*, **28**(1), 1-32. viewed 13 May 2020,

<<https://academic.oup.com/rfs/article/28/1/1/1682440>>

Dimpfl, T., and Jank, S. (2016). Can internet search queries help to predict stock market volatility?. *European Financial Management*, **22**(2), 171-192. viewed 24 February 2020,

<<https://onlinelibrary.wiley.com/doi/pdf/10.1111/eufm.12058>>

Egan, M. (2019). This app completely disrupted the trading industry, *CNN Business*, viewed 14 December 2019,

<<https://edition.cnn.com/2019/12/13/investing/robinhood-free-trading-fractional-shares/index.html>>

Ettredge, M., Gerdes, J., and Karuga, G. (2005). Using web-based search data to predict macroeconomic statistics. *Communications of the ACM*, **48**(11), 87-92. viewed 23 January 2020,

<<https://www.researchgate.net/publication/200110929-Using-Web-based-search-data-to-predict-macroeconomic-statistics>>

Feise, R. J. (2002). Do multiple outcome measures require p-value adjustment?. *BMC*

medical research methodology, **2**(1), 8. viewed 9 November 2020,

<<https://link.springer.com/article/10.1186/1471-2288-2-8>>

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, **457**(7232), 1012-1014. viewed 9 March 2020,

<<https://storage.googleapis.com/pub-tools-public-publication-data/pdf/34503.pdf>>

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., and Watts, D. J. (2010). Predicting consumer behavior with Web search, *Proceedings of the National academy of sciences*, **107**(41), 17486-17490. viewed 12 December 2019,

<<https://www.pnas.org/content/pnas/107/41/17486.full.pdf>>

Google. (2020). FAQ about Google Trends data. viewed 11 February 2020,

<<https://support.google.com/trends/answer/4365533?hl=en>>

Granell, A., and Carlsson, F. (2018). How Google Search Trends Can Be Used as Technical Indicators for the SP500-Index: A Time Series Analysis Using Granger's Causality Test. viewed 12 June 2020,

<<https://kth.diva-portal.org/smash/get/diva2:1210760/FULLTEXT01.pdf>>

Guzman, G. (2011). Internet search behavior as an economic forecasting tool: The case of inflation expectations. *Journal of economic and social measurement*, **36**(3), 119-167. viewed 21 July 2020,

<<http://datascienceassn.org/sites/default/files/Internet%20search%20behavior%20as%20an%20economic%20forecasting%20tool%20the%20case%20of%20inflation%20expectations.pdf>>

Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., and Smith, N. A. (2009). Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 272-280). viewed 22 September 2020,

<<https://www.aclweb.org/anthology/N09-1031.pdf>>

Kristoufek, L. (2013). Can Google Trends search queries contribute to risk diversification?. *Scientific reports*, **3**, 2713. viewed 21 February 2020,

<<https://www.nature.com/articles/srep02713>>

Lee, W. Y., Jiang, C. X., and Indro, D. C. (2002). Stock market volatility, excess returns, and the role of investor sentiment, *Journal of banking Finance*, **26** (12), 2277-2299. viewed 12 December 2019,

<<https://www.sciencedirect.com/science/article/pii/S0378426601002023>>

- Leinweber, D. J. (2007). Stupid data miner tricks: overfitting the SP 500. *The Journal of Investing*, **16**(1), 15-22. viewed 7 March 2020,
<https://www.researchgate.net/publication/247907373_Stupid_Data_Miner_Tricks_Overfitting_the_SP_500>
- Li, X., Ma, J., Wang, S., and Zhang, X. (2015). How does Google search affect trader positions and crude oil prices?. *Economic Modelling*, **49**, 162-171. viewed 20 February 2020,
<<https://www.sciencedirect.com/science/article/abs/pii/S0264999315001066>>
- Liu, B. (2012). Sentiment analysis and opinion mining, *Synthesis lectures on human language technologies*, **5** (1), 1-167. viewed 12 December 2019,
<<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.244.9480rep=rep1type=pdf>>
- Loughran, T., and McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, *The Journal of Finance*, **66**(1), 35-65. viewed 13 December 2019,
<<https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.2010.01625.x>>
- Massicotte, P. (2019). Package ‘gtrendsR’, viewed 2 December 2019,
<<https://cran.r-project.org/web/packages/gtrendsR/gtrendsR.pdf>>
- Nuti, S. V., Wayda, B., Ranasinghe, I., Wang, S., Dreyer, R. P., Chen, S. I., and Murugiah, K. (2014). The use of google trends in health care research: a systematic review. *PloS one*, **9**(10). viewed 6 April 2020,
<<https://journals.plos.org/plosone/article/file?type=printableid=10.1371/journal.pone.0109583>>
- Perlin, M. S., Caldeira, J. F., Santos, A. A., and Pontuschka, M. (2017). Can we predict the financial markets based on Google’s search queries?. *Journal of Forecasting*, **36**(4), 454-467. viewed 1 June 2020,
<<https://onlinelibrary.wiley.com/doi/pdf/10.1002/for.2446>>
- Polgreen, P. M., Chen, Y., Pennock, D. M., Nelson, F. D., and Weinstein, R. A. (2008). Using internet searches for influenza surveillance. *Clinical infectious diseases*, **47**(11), 1443-1448. viewed 9 January 2020,
<<https://academic.oup.com/cid/article/47/11/1443/282247>>
- Preis, T., Moat, H. S., and Stanley, H. E. (2013). Quantifying trading behavior in financial markets using Google Trends, *Scientific reports*, **3**, 1684. viewed 29 November 2019,
<https://www.researchgate.net/publication/236338265_Quantifying_Trading_Behavior_in_Financial_Markets_Using_Google_Trends>
- Provalis Research. (2018). WordStat Sentiment Dictionary 2.0, viewed 10 December 2019,

<<https://provalisresearch.com/products/content-analysis-software/wordstat-dictionary/sentiment-dictionaries/>>

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International journal of forecasting*, **16**(4), 437-450. viewed 29 December 2019, <<https://www.researchgate.net/publication/247087596-Out-of-sample-tests-of-forecasting-accuracy-a-tutorial-and-review>>

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, **62**(3), 1139-1168. <<https://onlinelibrary.wiley.com/doi/full/10.1111/j.1540-6261.2007.01232.x>>

U.S. Bureau of Labor Statistics. (2020). CPI Inflation Calculator. viewed 12 March 2020, <<https://data.bls.gov/cgi-bin/cpicalc.pl?cost1=1year1=201001year2=202001>>

Urquhart, A. (2018). What causes the attention of Bitcoin?. *Economics Letters*, **166**, 40-44. viewed 15 February 2020, <<https://www.sciencedirect.com/science/article/abs/pii/S016517651830065X>>

Vosen, S., and Schmidt, T. (2011). Forecasting private consumption: survey-based indicators vs. Google trends. *Journal of forecasting*, **30**(6), 565-578. viewed 9 January 2020, <<https://onlinelibrary.wiley.com/doi/full/10.1002/for.1213>>

Walimbe, R. (2017). Avoiding Look Ahead Bias in Time Series Modelling. *Date Science Central*. viewed 15 October 2020, <<https://www.datasciencecentral.com/profiles/blogs/avoiding-look-ahead-bias-in-time-series-modelling-1>>

Wilder, J. W. (1978). New concepts in technical trading systems, Trend Research. viewed 2 April 2019, <<http://agris.fao.org/agris-search/search.do?recordID=US201300554903>>

Wu, L., and Brynjolfsson, E. (2015). The future of prediction: How Google searches foreshadow housing prices and sales. In *Economic analysis of the digital economy* (pp. 89-118). University of Chicago Press. viewed 22 January 2020, <<https://www.nber.org/chapters/c12994>>

Xiong, R., Nichols, E. P., and Shen, Y. (2015). Deep learning stock volatility with google domestic trends. viewed 17 February 2020, <<https://arxiv.org/pdf/1512.04916.pdf>>

Yang, X., Pan, B., Evans, J. A., and Lv, B. (2015). Forecasting Chinese tourist volume

with search engine data. *Tourism Management*, **46**, 386-397. viewed 24 April 2020,
<<https://www.sciencedirect.com/science/article/abs/pii/S0261517714001514>>

Yen, S. M. F., and Hsu, Y. L. (2010). Profitability of technical analysis in financial and commodity futures markets - A reality check, *Decision Support Systems*, **50**(1), 128-139. viewed 5 April 2020,
<<https://www.sciencedirect.com/science/article/pii/S0167923610001144>>

Zhang, F., and Chen, J. Y. (2010). Discovery of pathway biomarkers from coupled proteomics and systems biology methods. *BMC genomics*, **11**(2), S12. viewed 3 April 2020,
<<https://bmcbgenomics.biomedcentral.com/articles/10.1186/1471-2164-11-S2-S12>>

Zhong, X., and Raghiv, M. (2019). Revisiting the use of web search data for stock market movements. *Scientific reports*, **9**(1), 1-8. viewed 22 February 2020,
<<https://www.nature.com/articles/s41598-019-50131-1>>

Zivot, E., and Wang, J. (2003). Rolling analysis of time series. *Springer* (pp. 299-346). New York, NY. viewed 11 March 2020,
<<https://link.springer.com/chapter/10.1007/978-0-387-21763-59>>

Appendix

The following table lists the transaction numbers needed to realize the first and second approach of the analysis of the collective signal strength in Section 6.2.7:

number of keywords	short-holding period (days)							
	98% sig. level	1	2	3	4	5	6	7
	1	530	290	162	126	78	50	42
	2	130	90	50	42	22	22	22
	3	10	10	6	6	6	6	6

number of keywords	short-holding period (days)							
	99.75% sig. level	1	2	3	4	5	6	7
	1	102	70	46	42	26	14	10
	2	2	2	2	2	2	2	2
	3	2	2	2	2	2	2	2

aggregated t-statistics	short-holding period (days)							
	98% sig. level	1	2	3	4	5	6	7
	2 - 3	422	222	122	98	66	54	50
	3 - 4	58	38	34	30	26	22	22
	4 - 5	142	110	78	66	54	46	42
	5 - 6	70	54	42	42	30	30	30
	6 - 7	2	2	2	2	2	2	2
	7 - 8	10	10	6	6	6	6	6

aggregated t-statistics	short-holding period (days)							
	99.75% sig. level	1	2	3	4	5	6	7
	2 - 3	2	2	2	2	2	2	2
	3 - 4	50	34	30	30	26	22	22
	4 - 5	50	42	34	30	30	26	22
	5 - 6	18	14	14	14	10	10	10
	6 - 7	2	2	2	2	2	2	2
	7 - 8	2	2	2	2	2	2	2

Figure 16: Analysis of the collective signal strength - transactions for the first and second approach

Source: Own figure

The following table lists all the keywords with their descriptive statistics of their search volumes:

keyword	n	start date	end date	mean	std dev	min	1st Q	median	3rd Q	max	ACF(1)	NAs
reforms	6085	2004-01-01	2020-08-28	44.71	28.40	0.0	23.0	39.3	64.4	160.9	0.51	0
boost	6085	2004-01-01	2020-08-28	71.48	15.44	0.0	62.8	73.1	81.6	155.9	0.65	0
consolidate	6085	2004-01-01	2020-08-28	55.05	23.34	0.0	40.1	55.3	71.1	244.5	0.39	0
construction	6085	2004-01-01	2020-08-28	75.84	19.14	0.0	54.9	83.6	91.1	109.3	0.46	0
outperform	6085	2004-01-01	2020-08-28	28.98	29.01	0.0	0.0	30.6	47.7	221.3	0.20	13
savings	6085	2004-01-01	2020-08-28	46.39	42.21	0.0	8.0	30.0	84.3	527.1	0.84	0
progress	6085	2004-01-01	2020-08-28	68.44	18.27	0.0	53.4	70.5	83.3	150.4	0.50	0
booming	6085	2004-01-01	2020-08-28	39.57	28.63	0.0	21.7	39.4	57.0	305.3	0.29	0
accrue	6085	2004-01-01	2020-08-28	39.11	29.28	0.0	18.4	38.5	58.2	255.8	0.34	2
surpass	6085	2004-01-01	2020-08-28	40.21	27.63	0.0	24.2	40.2	57.3	164.0	0.34	0
debt	6085	2004-01-01	2020-08-28	66.47	18.87	0.0	53.9	68.1	80.2	208.2	0.69	0
crisis	6085	2004-01-01	2020-08-28	62.15	20.07	0.0	47.6	60.3	76.2	190.8	0.74	0
decline	6085	2004-01-01	2020-08-28	61.18	20.85	0.0	46.3	60.1	76.1	200.0	0.51	0
recession	6085	2004-01-01	2020-08-28	47.30	26.96	0.0	30.0	45.3	62.9	790.2	0.53	0
unemployment	6085	2004-01-01	2020-08-28	58.75	20.01	0.0	45.6	58.4	71.4	285.0	0.36	0
bankrupt	6085	2004-01-01	2020-08-28	48.69	23.14	0.0	33.6	47.4	61.1	391.3	0.48	0
deficit	6085	2004-01-01	2020-08-28	52.59	20.26	0.0	38.2	50.7	64.4	243.8	0.59	0
collapse	6085	2004-01-01	2020-08-28	41.29	56.17	0.0	21.6	39.0	54.6	1950.0	0.18	0
market crash	6085	2004-01-01	2020-08-28	39.74	38.90	0.0	20.4	35.2	54.3	2027.4	0.29	0
downturn	6085	2004-01-01	2020-08-28	33.60	28.22	0.0	0.0	33.2	51.0	202.8	0.28	10
banana	6085	2004-01-01	2020-08-28	70.96	12.58	0.0	63.5	70.1	77.6	123.2	0.57	0
door	6085	2004-01-01	2020-08-28	82.31	10.15	0.0	77.7	83.5	88.5	143.0	0.67	0
weather	6085	2004-01-01	2020-08-28	57.97	17.50	0.0	46.0	57.5	68.0	159.7	0.84	0
chopstick	6085	2004-01-01	2020-08-28	40.24	26.71	0.0	24.7	39.9	55.8	195.9	0.31	0
phone	6085	2004-01-01	2020-08-28	84.79	10.07	0.0	78.2	85.9	92.5	126.6	0.56	0
chicago	6085	2004-01-01	2020-08-28	78.19	13.45	0.0	70.0	79.9	87.6	166.0	0.82	0
christmas	6085	2004-01-01	2020-08-28	61.69	39.57	0.0	14.2	75.6	91.7	195.9	0.94	104
swim	6085	2004-01-01	2020-08-28	67.12	21.66	0.0	49.5	70.4	83.3	189.4	0.88	0
bicycle	6085	2004-01-01	2020-08-28	43.60	24.81	0.0	27.7	41.1	57.0	190.2	0.22	0
programming	6085	2004-01-01	2020-08-28	77.48	11.22	0.0	69.6	77.9	85.5	113.4	0.43	0

Table 4: Descriptive statistics of the search volume of the keywords

Source: Own table

The following table lists the descriptive statistics of the log returns of the DJI and the sectoral indices:

ticker	n	start date	end date	mean	std dev	min	1st Q	median	3rd Q	max	ACF(1)	NAs
DJI	4192	2004-01-05	2020-08-27	0.02%	1.19%	-13.8%	-0.4%	0.0%	0.5%	10.8%	-14.35%	163
VNQ	4006	2004-09-30	2020-08-27	0.03%	1.97%	-21.7%	-0.6%	0.1%	0.8%	15.7%	-17.76%	0
XLK	4192	2004-01-05	2020-08-27	0.05%	1.34%	-14.9%	-0.5%	0.1%	0.7%	13.0%	-14.00%	0
XLB	4192	2004-01-05	2020-08-27	0.03%	1.52%	-13.3%	-0.6%	0.1%	0.8%	13.2%	-4.71%	0
XLV	4192	2004-01-05	2020-08-27	0.04%	1.08%	-10.4%	-0.4%	0.1%	0.6%	11.4%	-9.40%	0
XLI	4192	2004-01-05	2020-08-27	0.03%	1.35%	-12.0%	-0.5%	0.1%	0.7%	11.9%	-6.08%	0
XLP	4192	2004-01-05	2020-08-27	0.04%	0.89%	-9.9%	-0.4%	0.1%	0.5%	8.2%	-11.42%	0
XLE	4192	2004-01-05	2020-08-27	0.02%	1.89%	-22.5%	-0.8%	0.1%	0.9%	15.3%	-7.80%	0
XLY	4192	2004-01-05	2020-08-27	0.04%	1.33%	-13.5%	-0.5%	0.1%	0.7%	9.3%	-5.82%	0
VOX	4006	2004-09-30	2020-08-27	0.03%	1.28%	-12.1%	-0.5%	0.1%	0.6%	13.1%	-6.61%	0
XLU	4192	2004-01-05	2020-08-27	0.04%	1.18%	-12.1%	-0.5%	0.1%	0.6%	12.0%	-9.98%	0
XLF	4192	2004-01-05	2020-08-27	0.02%	2.00%	-18.2%	-0.6%	0.1%	0.7%	27.0%	-12.29%	0

Table 5: Descriptive statistics of the log returns of the DJI and the sectoral funds

Source: Own table