UNIVERSITY OF TURKU

ABSTRACT

	Bachelor's thesis
X	Master's thesis
	Licentiate's thesis
	Doctoral dissertation

Subject	Accounting and Finance	Date	30.5.2021
Author	Elmeri Saarenpää	Number of pages	74+appendices
Title	Explaining the cross-section of asset returns with Google Trends		
Supervisors	Prof. Markus Granlund, Dr. Jan Pfister		

Abstract

This thesis explains the cross-section of stock asset returns with Google Trends using methods from multiple studies. Risk factors set asset prices and explain differences in average returns. The five-factor model (FF5) of Fama and French (2015) can explain most of the expected stock returns. However, research on alternative factors may identify new risk premiums. Market sentiment combines behavioral finance with asset pricing. One approach quantifies sentiment from internet searches using Google Trends search volume indices. However, the methods of the previous studies may not apply to asset pricing.

In this study, I construct synthetic sentiment indices based on the Google Trends data consisting of the 98 search terms Preis, Moat, and Stanley (2013) propose. I use principal component analysis, suggested by Baker and Wurgler (2006), to construct the sentiment indices. Then, I explain the returns of both Fama–French portfolios and individual stocks with the sentiment indices. The models used in testing also contain the FF5 factors and self-constructed cross-sectional factors. Hypothesis testing includes various regression methods, such as the procedure of Fama and MacBeth (1973). The monthly data ranges from the year 2004 to 2017 and is limited to the United States.

The synthetic sentiment indices constructed in this thesis do not relate to aggregate stock market movements. Based on the *t*-statistics of regressions, the sentiment factors are not significant when applying appropriate clustering and standard error corrections methods. Also, the Google Trends-based sentiment does not explain the cross-section of asset returns. The *t*-statistics of the Fama–MacBeth regressions do not indicate any compensation from the exposure to the Google Trends-based sentiment factors. The risk premiums are not significant even at a 90% confidence level.

Based on the literature, new factors and approaches may strengthen the understanding of the expected asset returns. The availability and the amount of Google Trends data make it an appealing source for sentiment proxies. However, it may not be as beneficial of a tool as previous studies indicate. Different search term specifications, periods, and data frequencies offer an opportunity for further research.

Key words	Asset pricing, risk premium, market sentiment, big data, principal component

TURUN YLIOPISTO

TIIVISTELMÄ

	Kandidaatintutkielma
X	Pro gradu -tutkielma
	Lisensiaatintutkielma
	Väitöskirja

Oppiaine	Laskentatoimi ja rahoitus	Päivämäärä	30.5.2021
Tekijä	Elmeri Saarenpää	Sivumäärä	74+liitteet
Otsikko	Arvopaperien tuottojen poikkileikkauksen selitys Google Trends -datan avulla		
Ohjaajat	Prof. Markus Granlund, Dr. Jan Pfister		

Tiivistelmä

Tämä tutkielma selittää arvopaperien tuottojen poikkileikkausta Google Trends -datan avulla käyttäen eri tutkimusten menetelmiä. Riskifaktorit määrittävät arvopaperien hinnat ja selittävät keskimääräisiä tuottoeroja. Faman ja Frenchin (2015) viiden faktorin malli (FF5) selittää suurimman osan odotetuista osaketuotoista. Vaihtoehtoisten faktoreiden tutkimus saattaa kuitenkin tunnistaa uusia riskipreemioita. Markkinasentimentti yhdistää käyttäytymistieteellisen rahoituksen arvopaperien hinnoittelun. Yksi lähestymistapa mittaa sentimenttiä Google Trends -palvelun hakuvolyymi-indeksien avulla. Aiempien tutkimusten menetelmät saattavat kuitenkin olla soveltumattomia arvopapereiden hinnoitteluun.

Tässä tutkimuksessa Google Trends -dataan pohjautuvat synteettiset sentimentti-indeksit perustuvat 98 Preisin, Moatin and Stanleyn (2013) ehdottamaan hakusanaan. Sentimentti-indeksit muodostetaan käyttämällä pääkomponenttianalyysia, kuten Baker ja Wurgler (2016) suosittelevat. Sekä Fama–French-portfolioiden että yksittäisten osakkeiden tuottoja selitetään näiden sentimentti-indeksien avulla. Testauksessa käytettävät hinnoittelumallit sisältävät myös FF5-faktorit ja itse luodut poikkileikkausfaktorit. Hypoteesin testaukseen sovelletaan useita regressiomenetelmiä, kuten Faman ja MacBethin (1973) menetelmää. Kuukausidata vuosilta 2004–2017 on rajattu Yhdysvaltoihin.

Tässä tutkielmassa luoduilla synteettisillä sentimentti-indekseillä ei ole yhteyttä osaketuottoihin kokonaisuudessaan. Regressiotulosten *t*-arvojen mukaan sentimentti ei ole merkitsevä asiaankuuluvan klusteroinnin ja keskivirheiden korjauksen jälkeen. Google Trends -dataan pohjautuva sentimentti ei myöskään selitä arvopapereiden tuottojen poikkileikkausta. Fama–MacBeth-regressioiden *t*-arvot osoittavat, ettei altistumisesta sentimentti-indekseille palkita. Riskipreemiot eivät ole merkitseviä edes 90 % luottamustasolla.

Kirjallisuus esittää uusien faktoreiden ja menetelmien voivan vahvistaa käsitystä arvopapereiden odotetuista tuotoista. Google Trends -datan saatavuus ja laajuus tekevät siitä houkuttelevan lähteen sentimenttianalyysiin, mutta tämän hyödynnettävyys saattaa olla uskottua heikompi. Erilaiset hakutermiyhdistelmät, aikaperiodit ja datafrekvenssin muuttaminen tarjoavat jatkotutkimusmahdollisuuksia.

Avainsanat	Arvopaperien hinnoittelu, riskipreemio, sentimentti, big data, pääkomponentti

UNIVERSITY OF TURKU
Turku School of Economics

EXPLAINING THE CROSS-SECTION OF ASSET RETURNS WITH GOOGLE TRENDS

Master's Thesis
in Accounting and Finance

Author:
Elmeri Saarenpää

Supervisors:
Prof. Markus Granlund
Dr. Jan Pfister

30.5.2021
Turku

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.

CONTENTS

FIGURES

TABLES

1 INTRODUCTION

1.1 Background and motivation

The relationship between risk and return sets asset prices. Consequently, the asset's price should equal its expected discounted payoff. The discount factor is the investor's marginal utility-price ratio. The investor's ratio of marginal utilities is the fundamental, but theoretical, discount factor (Cochrane 2005, 3–4). The consumption-based capital asset pricing model (CCAPM) of Lucas Jr (1978) and Breeden (1979) measures the marginal utility through consumption. Most of the asset pricing theory focuses on different variations of this model. The CCPAM is the most integral asset pricing model from a theoretical standpoint. However, it does not work well in practice (Cochrane 2005, 6, 41).

Factor pricing models proxy the marginal utility with a linear model. The capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965) is a single-factor model relying on market beta to explain the expected return of an asset. The capital asset pricing model has more practical implications than the consumption-based model. However, it still relies on strict assumptions and has genuine real-world limitations (Black, Jensen and Scholes 1972). The arbitrage pricing theory (APT) of Ross (1976) is a multi-factor approach to asset pricing. The linear relationship between various variables capturing systematic risk explains the asset returns. The APT requires less restrictive assumptions than the CAPM, but it does not specify which factors to use (Ross 1976).

Fama and French (1993) introduce the three-factor model (FF3) with size risk and value factors based on the findings of Banz (1981) and Basu (1983). Later, Fama and French (2015) add the profitability and investment factors to the FF3 model to create the Fama–French five-factor model (FF5). The FF5 is a more theoretically sound model than the FF3, which is an empirical model. Fama and French (2015) derive the factors from the dividend discount model and the dividend policy irrelevance of Miller and Modigliani (1961). Subrahmanyam (2010) identifies more than fifty factors explaining the cross-section of stock returns. However, finance theory generally accepts only a handful of them as independent risk factors. The FF5 can explain most of the expected stock returns. However, research on alternative factors may identify new risk premiums.

Fama and MacBeth (1973) introduce a method to evaluate asset pricing models. The model can also be used to generate standard errors correcting for cross-sectional correlation. The Fama–MacBeth (FMB) procedure consists of two steps

to estimate betas and risk premiums. It uses a rolling estimation procedure during the first step. The FMB is the most used method to correct standard errors in finance studies, based on Petersen (2009). Although more modern methods exist, the FMB remains commonly used in academic literature (Cochrane 2005, 245–252).

The changes in risk premiums can explain the majority of long-term aggregate asset returns. However, they may fail to do so in the short term. Market sentiment refers to the overall attitude towards the anticipated price development of the market. The market sentiment can be quantified and used to explain asset returns (Barberis, Shleifer and Vishny 1998). Therefore, market sentiment combines aspects of behavioral finance with asset pricing. Different approaches to quantify investor attention include financial markets, non-financial factors, surveys, text mining, and internet search behavior-related measures (see, e.g., Barberis et al. 1998; Baker and Wurgler 2006; 2007). Although the market sentiment is an increasingly popular subject in academic studies, the classic finance theory does not recognize it (Baker and Wurgler 2006).

One popular approach uses Google Trends search volume data to capture investor attention. The availability and the amount of Google Trends data make it an appealing source for sentiment proxies (Da, Engelberg and Gao 2011). Preis, Moat and Stanley (2013) study 98 different search terms and create a trading strategy based on their search occurrence. Consequently, trading based on the keyword *debt*, for example, outperforms the buy-and-hold strategy significantly. Preis et al. (2013) measure the search term's financial relevance by the Financial Times appearances. The keyword-relevance has a significant positive correlation with its likelihood to forecast subsequent stock market moves (Preis et al. 2013).

Baker and Wurgler (2006) study how the market sentiment explains the cross-section of asset returns. No single perfect proxy exists, so they form a sentiment index from multiple measures using principal component analysis (PCA). PCA is a multivariate analysis technique for finding patterns in large data sets (Abdi and Williams 2010). Da, Engelberg and Gao (2015) combine Google Trends data with PCA. They use PCA to form a sentiment index from negatively associated keywords and find it to predict price reversals, volatility spikes, and fund flows.

1.2 Research questions and objectives

This study aims to explain the cross-sections of asset returns with Google Trends. It examines if the synthetic sentiment indices constructed in the thesis relate to aggregate stock market movements. It also analyzes if the Google Trends-based sentiment explains the cross-section of asset returns. The market sentiment proxy studies, such as Google Trends data, show promising results. However, those studies mostly focus on sentiment as a trading strategy. Therefore, their outcomes and methodologies may not apply to asset pricing. When tested with appropriate methods, the Google Trends-based sentiment may not be as useful of a tool as previous studies indicate.

I combine methods from multiple studies to test the relevance of aggregate Google Trends search volume data in asset pricing. I use the Google search terms Preis et al. (2013) propose to see if they contain relevant information when tested with a more appropriate methodology. I use PCA to construct the synthetic sentiment indices as suggested by Baker and Wurgler (2006). One of these sentiment factors uses the whole search term sample, and the other factors use a smaller search term selection. I explain the returns of both Fama–French portfolios and individual stocks with these sentiment factors. The models I use in testing also contain the five of the Fama and French (2015) factors and self-constructed cross-sectional factors. Hypothesis testing includes various regression methods, such as the procedure of Fama and MacBeth (1973). Other methods include regressions with clustering suggested by Cochrane (2005, 245–252) and Petersen (2009), so-called long-short regressions from Baker and Wurgler (2006), and risk-adjusted regressions inspired by Brennan, Chordia and Subrahmanyam (1998).

The two research questions of this thesis are the following:

1. Are the synthetic sentiment indices constructed in the thesis related to aggregate stock market movements?

2. Does the Google Trends-based sentiment explain the cross-section of asset returns?

The first research question examines if aggregate Google search volume changes explain or anticipate movements in aggregate stock returns. I answer the question by testing the coefficients of the Google Trends-based sentiment factors. If they are significantly different from zero when I apply appropriate regression methods and adjustments, I reject the null hypothesis. In this case, the sentiment indices relate to aggregate stock market movements. The second research question investigates

if the sentiment can explain the cross-section of expected returns. I respond to this question by testing the lambda coefficients of the sentiment indices. I assess this research question mostly with the FMB regressions. Consequently, if the lambda coefficients significantly differ from zero, I reject the null hypothesis. In this case, the Google Trends-based sentiment explains the cross-section of asset returns.

1.3 Anticipated contributions and limitations

This thesis contributes to asset pricing literature and especially to a particular segment studying the explanatory capabilities of alternative factors in multi-factor models. In the arbitrage pricing theory of Ross (1976), a linear combination of systematic risk factors prices assets. Traditionally, these factors, such as size and value, are derived from firm characteristics. However, these factors can be anything believed to proxy marginal utility growth. Market sentiment, or investor attention, is an increasingly popular subject in academic studies combining behavioral finance and asset pricing. It explores different approaches to quantify investor attention, including internet search behavior measures (see, e.g., Barberis et al. 1998; Baker and Wurgler 2006; 2007). In this thesis, I focus on a narrow niche examining Google Trends data as a proxy for market sentiment. The most notable studies within this field include Da et al. (2011; 2015). Nevertheless, there is both theoretical and practical gap in the literature regarding the use of Google Trends data in asset pricing.

The thesis also contributes to Google Trends-related research, in general, as it evaluates the usefulness of Google Trends data in theoretical applications. As above-mentioned, many related studies do not use asset pricing methodology. For example, Preis et al. (2013) focus on trading strategies and not on the cross-section of asset returns. They report impressive results, but these findings may apply to asset pricing. Also, the methodology leaves room for improvement. According to Chordia, Goyal and Saretto (2017), many studies fail to test the hypothesis correctly. Therefore, few trading strategies outperform the market. This methodological ambiguity creates a need to replicate the research of Preis et al. (2013) using the same search terms but in a more appropriate empirical setting.

Since the subject is recent and scarcely researched, this thesis aims to construct a naive framework for future related studies. I combine leading practices from multiple sources, such as the principal component analysis (PCA) approach suggested by Baker and Wurgler (2006). They use PCA to construct sentiment

indices from numerous financial measures. However, this study applies PCA to Google Trends data. You can use this sentiment index in an asset pricing model, such as the five-factor model of Fama and French (2015), which asset pricing literature considers a leading practice. The regression methods of this thesis include the procedure of Fama and MacBeth (1973), a frequently used tool in asset pricing studies, along with multiple alternative methods. This thesis also contributes to the respective literature and practical implications of these methods.

There is also a controversy in the literature since the classic finance theory does not recognize the alternative factors (Baker and Wurgler 2006). The Fama–French framework, which covers, e.g., three and five-factor models, is in line with the efficient market hypothesis (Fama and French 1993; 2015). Non-conventional factors do not fit into this framework, and therefore they are not independent risk factors. This thesis aims to evaluate the role of these factors in asset pricing, both from a theoretical and practical perspective. Like all other asset pricing studies, this thesis also contributes to the tests of market efficiency. I do not cover technical specifications of concepts, such as the principal component analysis, in great technical detail since they are not the focus of this study.

The empirical part of this study focuses on stocks instead of other assets. Geographically, I limit both Google data and financial data to the United States. Preis, Moat, Stanley and Bishop (2012) study Google search usage and find internet search behavior being more future-orientated in countries with a high gross domestic product, such as the United States. Also, the English language makes testing different search terms easier. Besides, the U.S. stock market is the largest globally. People have broad access to the internet, and they likely invest in their domestic market.

Google Trends data consists of the 98 search terms proposed by Preis et al. (2013), and the data source is limited to web searches. The period ranges between January 2004–December 2017 and is lengthier than in many previous studies. I choose this period since Google Trends data is only available from 2004 onward. Google Trends only provides monthly data for such a lengthy period (Stephens-Davidowitz and Varian 2014). This data frequency should be sufficient since empirical asset pricing often uses monthly frequency. Also, monthly data contains less statistical noise.

1.4 Structure of the study

Section 2 lays out the literature review and hypothesis. I derive asset pricing models from the basic pricing equation and present them as variations of the consumption-based capital asset pricing model. I present findings and criticism related to different asset pricing models and factors. Then, I demonstrate the FMB procedure, and it can be how it can test asset pricing models. Afterward, I cover literature related to market sentiment and focus on studies using Google Trends as a proxy for market sentiment. Finally, I present PCA and cover its applications in asset pricing and market sentiment studies.

Section 3 covers the data used in the empirical part of this study. I present the individual company data which I acquire from Thomson Reuters Datastream. I cover the procedures to clean the data and how I form the self-made factors from firm characteristics. Then I describe the Fama–French portfolio and factor data and a way to obtain it from the Data Library of French (2018). I present the Google Trends data and how I acquired it using statistical program R. Then, I explain how I construct the sentiment indices. Finally, I offer the regression methods.

Section 4 presents the empirical results. I demonstrate the sample descriptive statistics, including financial data of portfolios and individual companies. I also establish Google Trends data and the constructed sentiment indices. Then I present the FMB regression results on both single socks and Fama–French portfolios. Finally, I present the results from all other regression methods.

Section 5 provides a summary and conclusion of this thesis. First, I summarize the literature review and its main findings. Then, I draw conclusions based on empirical results and previous studies and give suggestions for further research. I present the R code in Appendix 1. It contains all the analyses I use in the empirical part of the study. You can download the Google Trends data with the script I provide in the code.

2 LITERATURE REVIEW AND HYPOTHESIS

2.1 Equilibrium models

An asset's price should equal its expected discounted payoff (Cochrane 2005, 3). Two different approaches to this concept lead to either absolute pricing or relative pricing. In absolute pricing or equilibrium pricing, the prices reflect exposure to macroeconomic risk. Relative pricing or risk-neutral pricing values an asset based on the values of other assets. The absolute pricing approach includes various factor models, whereas relative pricing leads to option pricing (Cochrane 2005, 183). The basic pricing equation of asset pricing is

$$p_t = \mathbb{E}(m_{t+1}x_{t+1}), \tag{1}$$

where p_t is asset price, x_{t+1} is asset payoff, and m_{t+1} is a stochastic risk factor. Term m_{t+1} is a function of different data and parameters depending on the application and asset class, making Equation (1) both simple and universal. Term m_{t+1} is the investor's marginal utility-price ratio. The investor's ratio of marginal utilities is the fundamental measure asset pricing models observe. You can also express Equation (1) as $p = \mathbb{E}(mx)$ with subscripts suppressed (Cochrane 2005, 3, 6–7).

The consumption-based capital asset pricing model (CCAPM) of Lucas Jr (1978) and Breeden (1979) measures the marginal utility through consumption. Accordingly, people maximize utility from consumption instead of wealth. The marginal utility loss of consuming less today should equal the marginal utility gain of consuming more in the future (Cochrane 2005, 3). The CCAPM uses a utility function defined over the current and future value of consumption

$$U(c_t, c_{t+1}) = u(c_t) + \beta_c \mathbb{E}_t[u(c_{t+1})],$$

where c_t stands for consumption at time t. Decreasing consumption to hold more assets leads to a margin utility loss. This loss should equal the marginal utility gain of increased consumption enabled by the asset's future payoff. Solving this utility maximization trade-off yields

$$p_t = \mathbb{E}_t\left[\beta_c \frac{u'(c_{t+1})}{u'(c_t)} x_{t+1}\right], \tag{2}$$

where the stochastic discount factor of Equation (1) expresses optimal consumption

choice

$$m_{t+1} \equiv \beta_c \frac{u'(c_{t+1})}{u'(c_t)}.$$

According to Cochrane (2005, 6, 41), Equation (2) is the central asset pricing formula. Most of the asset pricing theory focuses on its different variations. The CCPAM should be, at least in principle, a complete answer to all asset pricing questions, but it does not work well in practice. When tested with actual consumption data, Campbell and Cochrane (1999, 248) find the consumption-based model to capture long-term stock price movements. However, the capital asset pricing model (CAPM) and its various multi-factor extensions perform better in an empirical setting. Also, Mankiw and Shapiro (1986) find no empirical evidence to support CCAPM over other asset pricing models. The CCAPM functions the best as a theoretical model combining macroeconomics and asset pricing.

You can derive most of the asset pricing models through Equation (1). For stocks, the payoff x_{t+1} is the sum of price p_{t+1} and divided d_{t+1}. You can obtain gross returns can then dividing with price p_t

$$R_{t+1} \equiv \frac{x_{t+1}}{p_t}.$$

You can also consider returns as units of consumption and apply Equation (1) to returns instead of prices. Think of a return as the payoff with the price of one

$$1 = \mathbb{E}(mR). \tag{3}$$

The risk-free rate is specific, so Equation (1) can be

$$1 = \mathbb{E}(m)R_f,$$

where R_f is the risk-free rate, defined as

$$R_f = \frac{1}{\mathbb{E}(m)}. \tag{4}$$

Given the definition of covariance

$$cov(m, x) = \mathbb{E}(mx) - \mathbb{E}(m)\mathbb{E}(x),$$

You can rewrite Equation (1) as

$$p = \mathbb{E}(m)\mathbb{E}(x) + cov(m, x),$$

where p is the price. As shown in Equation (3), the price can be a payoff of one.

Therefore,

$$1 = \mathbb{E}(m)\mathbb{E}(R_i) + cov(m, R_i),$$

where i is an asset and 1 is the payoff. Applying $\mathbb{E}(m) = 1/R_f$ based on Equation (4) expresses the return of an asset i as

$$\mathbb{E}(R_i) = R_f - R_f cov(m, R_i). \tag{5}$$

If $cov(m, R_i) = 0$, then $\mathbb{E}(R_i) = R_f$. Therefore, an investor does not receive compensation from idiosyncratic risk. Only systematic risk generates a premium. You can express Equation (5) for asset i with a beta pricing model

$$\mathbb{E}(R_i) = R_f + \left(\frac{cov(R_{i,m})}{var(m)}\right)\left(-\frac{var(m)}{\mathbb{E}(m)}\right)$$

or

$$\mathbb{E}(R_i) = R_f + \beta_{i,m}\lambda_m, \tag{6}$$

where $\beta_{i,m}$ is the beta coefficient. It measures the asset-specific quantity of risk. Interpret the coefficient λ_m as a price of risk for all assets (Cochrane 2005, 8–16). Factor pricing models proxy the marginal utility growth with a linear model

$$m_{t+1} = a + b' f_{t+1},$$

where a and b are some parameters. The factor pricing aims to find variables proxying for marginal utility growth, and

$$\beta_c \frac{u'(c_{t+1})}{u'(c_t)} \approx a + b' f_{t+1} \tag{7}$$

is an appropriate approximation. The relation of these variables or factors should be linear (Cochrane 2005, 149).

The capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965) is one of the most well-known asset pricing models. It uses market return as a variable to proxy for marginal utility growth described in Equation (7). The CAPM is a single-factor variation of Equation (6), which relies on market beta to explain an asset's expected return. The expected return of an asset or a stock i can be

$$\mathbb{E}(R_i) = R_f + \beta_i[\mathbb{E}(R_m) - R_f], \tag{8}$$

where $\mathbb{E}(R_m)$ is the expected market return and term $\mathbb{E}(R_m) - R_f$ implies market risk premium. Consequently, $\mathbb{E}(R_i) - R_f$ refers to the risk premium of asset i. Beta β_i is the covariance of the return of an asset i with the market return m divided by the market return variance. The CAPM comes from the modern

portfolio theory (MPT) of Markowitz (1952) and its optimal investor behavior and portfolio selection. A rational investor would hold a portfolio on a capital market line (CML) based on their risk preferences measured by an indifference curve. CML is the tangent line formed by a risk-free asset and a market portfolio. The market portfolio is always efficient and contains no asset-specific or idiosyncratic risk (Tobin 1958). The Capital asset pricing model derives an individual asset's expected return through the risk-free rate and the expected market return. Beta measures the sensitivity of an asset concerning the overall market. Beta over one indicates an above-average risk, which results in a higher expected return due to higher systematic risk exposure and vice versa. You can also interpret Equation (8) as a security market line (SML) or a tangent representing the expected return of an asset i (Sharpe 1964).

The CAPM is an intuitive model, but it relies on strict assumptions. Therefore the model has genuine real-world limitations. Accordingly, all investors are rational, risk-averse, maximize their wealth, and can borrow or lend any amount at the risk-fee rate, for example (Black 1972). The only testable hypothesis around the CAPM is whether the market portfolio is mean-variance efficient. This hypothesis means no portfolio has lower risk and higher return. However, no portfolio is mean-variance efficient in practice (Gibbons, Ross and Shanken 1989). Roll (1977) claims the market portfolio to be unobservable because you must evaluate it against all possible investment opportunities. Based on these findings, the theoretical framework around the CAPM may not hold.

The CAPM also fails to explain the expected returns in empirical tests (Black et al. 1972). Banz (1981) criticizes the CAPM and demonstrates how small stocks have consistently higher average returns than can be explained by their market beta. Rosenberg, Reid and Lanstein (1985) present similar findings regarding value socks and how the market beta does not fully explain the average returns of stocks with high book-to-market ratios. All in all, the market beta appears to explain approximately two-thirds of the returns of a diversified portfolio.

You can test the efficient market hypothesis (EMH) of Fama (1970) with asset pricing models. According to the EMH, investors behave rationally, and the stock prices reflect all the available information at any given moment. These are, however, problematic assumptions to test empirically. Tests of market efficiency are jointly testing of asset pricing models and jointly tests of the EMH. This issue is known as the joint hypothesis problem. Based on the failure to explain the stock return, either the market is inefficient, or the asset pricing model is incorrect. An asset pricing model can always be flawed, and therefore, the EMH becomes a paradox (Campbell, Lo and MacKinlay 1997). The EMH cannot be confirmed ei-

ther unless an asset pricing model thoroughly explains the expected stock returns. However, the efficiency level is still measurable to a certain extent (Campbell et al. 1997, 24–25).

2.2 Arbitrage pricing theory

The basic equilibrium models rely on a single beta. In contrast, the arbitrage pricing theory (APT) of Ross (1976) is a multi-factor approach to asset pricing. The linear relationship between various variables capturing systematic risk explains the asset returns. The expected return of an asset i is

$$\mathbb{E}(R_i) = R_f + \sum_{k=1}^{n} \beta_{ik} x_k,$$

where x_k is an unknown factor. The APT derives prices from arbitrage arguments and assumes the market can occasionally price assets incorrectly. You can then capture this pricing inefficiency through various additional factors. The APT requires less restrictive assumptions than the CAPM, but it does not specify which factors to use (Ross 1976).

Researchers have identified many variables to capture systematic risk. One of the first identified factors is based on the size effect. Accordingly, small companies have higher risk-adjusted returns and a nonlinear relation to the market factor (Banz 1981). Basu (1983) find similar evidence of the value effect measured by earnings' yield or book-to-market value. On average, stocks with high earnings yield have higher risk-adjusted returns. The effect persists regardless of the firm size. Rosenberg et al. (1985) provide further evidence of the value effect measured by the book-to-market ratio. When used in a multi-factor approach with the excess market return factor, the size and value factors can effectively capture the average stock returns (Fama and French 1992).

Fama and French (1993) create their three-factor model (FF3) based on earlier findings. The FF3 adds two well-established factors, size risk and value premium, to the beta pricing model. The excess return for asset i can be expressed as

$$\mathbb{E}(R_i) = R_f + \beta_{i,(R_m - R_f)}(R_m - R_f) + \beta_{i,SMB} SMB + \beta_{i,HML} HML, \tag{9}$$

where *SMB* or small minus big is the excess return of stocks with a small market capitalization less the stocks with large market capitalization. Also, *HML* or high minus low is the returns of stocks with a high book-to-market ratio minus

the returns of stocks with a low book-to-market ratio (Fama and French 1993). Accordingly, the stocks with small and value characteristics contain more risk and generate higher returns than the market or stocks with the opposite characteristics. If the market efficiency applies, investors require a size premium for small stocks with higher business risk and higher cost of capital. Value stocks suffer from dire outlooks, poor earnings, low profitability, and questionable financial strength (see, e.g., Fama and French 1993; Chen and Zhang 1998). The explanations apply with the efficient market hypothesis (see, e.g., Fama 1970; Fama and French 1993). Accordingly, the size and value premiums are as evident as the existence of the market premium. The FF3 model explains about 90 percent of a diversified portfolio's returns (Fama and French 1993).

The factor selection and results of Fama and French (1993) also raise criticism. Lo and MacKinlay (1990a) suggest data dredging as an explanation for the size and value effect. Petkova (2006) argues a model incorporating a set of macroeconomic variables, such as term and credit spreads, should outperform the FF3's ability to explain the cross-section of asset returns. Basu (1983) finds the size effect to disappear when controlling with earnings' yield. Therefore, the size can be beneficial when used with other factors to predict excess returns. However, it may not be robust as an independent factor (Basu 1983). Factor effects, in general, appear to be stronger among small stocks. Consequently, growth stocks with low market capitalization fail to deliver the size effect, lowering the factor's significance (Fama and French 1993; 2015).

However, even the most established factors, such as value, may experience long periods of weak performance and still be considered significant (Fama and French 2020). Fama and French (2020) discuss the decline or disappearance of the value premium following the Fama and French (1992; 1993). Although the value premium is now weaker, it is still significant in the whole observable period due to the strong performance before 1992. Fama and French (2020) believe long periods of varying performance to be natural and the value premium to be significant. Accordingly, the average performance may be the best trajectory for the expected value premium in the future. The high volume of monthly premiums also supports the value premium's persistence (Fama and French 2020).

Jegadeesh and Titman (1993) discover the momentum effect by studying cross-sectional momentum strategies. They buy recently outperformed stocks and sell underperforming stocks. Accordingly, stocks' relative performance tends to continue in the near-term future, but they cannot fully explain why. The momentum effect's persistence is one of the biggest arguments against the efficient market hypothesis (Fama 1970). It does not have a sensible explanation (Jegadeesh and

Titman 1993). Jegadeesh and Titman (2001) find the effect to persist after its discovery. Therefore, the initial results may not be a result of data dredging. They also suggest behavioral models as the best explanation for the effect.

Carhart (1997) studies mutual funds' performance and expands the FF3 model described in Equation (9) with the momentum factor. This expansion creates the Carhart four-factor model

$$\begin{aligned}\mathbb{E}(R_i) = R_f + \beta_{i,(R_m-R_f)}(R_m - R_f) + \beta_{i,SMB}SMB \\ + \beta_{i,HML}HML + \beta_{i,MOM}MOM,\end{aligned} \tag{10}$$

where MOM is the momentum factor based on the past 12-month performance. Accordingly, the additional momentum factor in the model can explain the under or over-performance of mutual funds. Therefore, an actively managed fund's performance may not reflect the fund manager's skill (Carhart 1997).

Because the momentum effect contradicts the EMH, it is problematic for the Fama–French framework. Momentum is nearly opposite to the value factor, which also causes problems for the framework. According to Fama and French (2006), profitability and investment have an impact on expected stock returns. Accordingly, more profitable companies and companies which invest conservatively tend to have higher average returns. According to Novy-Marx (2013), a multi-factor model using the profitability factor can effectively explain most of the earnings-related anomalies. Aharoni, Grundy and Zeng (2013) find similar, although slightly weaker, results about the expected investment.

Fama and French (2015) add the profitability and investment factors to the FF3. The model captures the momentum effect and complies with the EMH. It is known as the Fama–French five-factor model (FF5)

$$\begin{aligned}\mathbb{E}(R_i) = R_f + \beta_{i,(R_m-R_f)}(R_m - R_f) + \beta_{i,SMB}SMB \\ + \beta_{i,HML}HML + \beta_{i,RMW}RMW + \beta_{i,CMA}CMA,\end{aligned} \tag{11}$$

where profitability factor RMW stands for the robust minus week. Investment factor CMA stands for the conservative minus aggressive. The FF3 is an empirical model explaining the most basic anomalies. However, it does not have a comprehensive theoretical background (Fama and French 2015). The FF5 model is more theoretically sound. Fama and French (2015) derive the factors from the dividend discount model and the dividend policy irrelevance of Miller and Modigliani (1961). The five factors combined can explain most of the expected stock returns. Although the value factor itself may be significant, it is relatively insignificant when controlled with the market beta, size, profitability, and investment factor (Fama

and French 2015). The FF5 model still struggles to explain small stocks' returns with high investment and low profitability (Fama and French 2015). Nowadays, the FF5 is the leading practice in empirical asset pricing.

More than fifty factors explaining the cross-section of stock returns are identified and studied. However, finance theory only recognizes a handful of them as independent risk factors. These include beta, size, book-to-market, momentum, profitability, and investment, for example (Subrahmanyam 2010). Many factors face mixed results despite being studied comprehensively. Well-known factors often explain the results of new factors. Miller and Scholes (1982) study dividend yield concerning the expected stock returns but do not find enough evidence to support the factor. Dividend yield may be related to the long-term expected returns. However, the effect can be explained mostly by temporary shocks in current prices (Fama and French 1988). Fama and French (1992) study leverage a predictor and find a book-to-market factor to mainly explain the impact.

Sloan (1996) studies accrual and cash flow components of current earnings and finds the prices not fully reflecting this information. Accordingly, higher accruals and asset growth, in general, indicate subsequent low returns. Fama and French (2008) support these findings. Piotroski (2000) studies accounting fundamentals-based strategies among stocks with high book-to-market ratios. He aims to identify future winners from losers based on these fundamentals. Consequently, a portfolio of financially sound companies outperforms their peers. The results indicate the market may not fully account for historical financial information (Piotroski 2000). The stock issuance and repurchase may also predict stock returns (Fama and French 2008).

Chen, Roll and Ross (1986) find evidence of macroeconomic variables affecting stock market returns systematically. These include interest rate spreads, inflation, industrial production, and the bond yield spreads. However, the oil price is not a significant variable, unlike intuition might suggest. According to the empirical evidence, changes in consumption do not significantly affect asset returns, which contradicts the CCAPM (Chen et al. 1986). Ang, Hodrick, Xing and Zhang (2006) suggest that past volatility negatively impacts expected stock returns.

Increasingly popular sustainable investing based on environmental, social and governance (ESG) criteria can be considered factor investing. Pastor, Stambaugh and Taylor (2019) find a connection between ESG preferences and asset prices. Investors with strong ESG preferences earn lower expected returns but receive utility from holding more sustainable assets. This utility gain should exceed the loss in expected return. Consequently, higher expected returns compensate for exposure to ESG risk. Consequently, returns of non-ESG portfolios are more

vulnerable to unexpected ESG concerns. Polarized ESG preferences also highlight this effect (Pastor et al. 2019).

2.3 Fama–MacBeth procedure

Fama and MacBeth (1973) suggest an alternative cross-sectional regression method using a rolling estimation procedure to estimate asset pricing models' parameters. The Fama–MacBeth (FMB) procedure runs cross-sectional regressions and generates standard errors correcting cross-sectional correlations. The method consists of two steps to estimate betas and risk premium for factors. The model is beneficial in testing asset pricing models and can be used in pooled and panel regressions. More modern methodologies, such as clustering methods, also exist. However, the FMB is relatively easy to implement and remains commonly used in academic literature and empirical finance (Cochrane 2005, 245–252).

The first step of the FMB procedure estimates factor loadings. This step runs linear time series regressions

$$R_{i,t} - R_{f,t} = \alpha_i + \beta_i x_t + \epsilon_{i,t}, \tag{12}$$

where x_t represents the factors used to explain the excess return of an asset i. The linear time series regression has multiple variations. Fama and MacBeth (1973) use a 5-year rolling regression during the first step. Brennan et al. (1998) use a somewhat similar rolling method in their Fama–MacBeth procedure, whereas Petkova (2006) also performs complete sample estimates. The results may differ depending on the method and length of the rolling period. Beta estimator values minimize the sum of squared residuals

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x-\bar{x})^2(y-\bar{y})^2}{\sum_{i=1}^{n}(x-\bar{x})^2} = \frac{s_{x,y}}{s_x^2}, \tag{13}$$

where $\bar{x}$ and $\bar{y}$ are the averages of x and y. Term $s_{x,y}$ is the sample covariance, and s_x^2 is the sample variance. The term x_t in Equation (12) can be defined, for example, as the FF5 model factors

$$x_t \equiv [(R_{m,t} - R_{f,t}), SMB_t, HML_t, CMA_t, RMW_t]. \tag{14}$$

After estimating risk exposure to betas, the second step estimates factor risk premium by running cross-sectional regressions for each time observation t. The

beta estimates are independent variables in the regression model

$$R_{i,t} - R_{f,t} = \alpha_i + \lambda_t \hat{\beta}_i + \epsilon_{i,t}, \tag{15}$$

where λ_t is lambda representing a factor risk premium at time t. Term $\hat{\beta}_i$ represents the beta estimators of risk factors for asset i. Respectively to Equation (14), the beta estimators are

$$\hat{\beta}_i \equiv [\hat{\beta}_{i,(R_m - R_f)}, \hat{\beta}_{i,SMB}, \hat{\beta}_{i,HML}, \hat{\beta}_{i,CMA}, \hat{\beta}_{i,RMW}]. \tag{16}$$

Once you have performed these two steps, the market price of factor risk is the average of individual factor risk premiums. In other words,

$$\hat{\lambda}_{FMB} = \frac{1}{T} \sum_{t=1}^{T} \hat{\lambda}_t, \tag{17}$$

where $\hat{\lambda}_{FMB}$ is the average of lambdas or the Fama–MacBeth risk premium estimates. As in Equation (14) and Equation (16), the lambda estimators for the FF5 factors can be

$$\hat{\lambda}_t \equiv [\hat{\lambda}_{t,(R_m - R_f)}, \hat{\lambda}_{t,SMB}, \hat{\lambda}_{t,HML}, \hat{\lambda}_{t,CMA}, \hat{\lambda}_{t,RMW}],$$

where $\hat{\lambda}_t$ are the lambda estimates. The FMB assumes the realizations of lambda estimates are approximately independent and identically distributed. If the betas are constant over time, then the FMB estimates are entirely identical to cross-sectional estimates (Cochrane 2005, 245–252).

The FMB standard errors, which correct for cross-sectional correlation, can be obtained as

$$SE_{\hat{\lambda}_t} = \frac{\sigma_{\hat{\lambda}_t}}{\sqrt{T}},$$

where $\sigma_{\hat{\lambda}_t}$ is the variance of the lambdas and T is the number of periods. The statistical significance can be determined with t-statistics to test whether the FMB risk premium estimates are statistically different from zero. You can obtain the t-statistics as

$$t_\lambda = \frac{\hat{\lambda}_{FMB}}{\sigma^2_{\hat{\lambda}_t} / \sqrt{T}}, \tag{18}$$

where $\sigma^2_{\hat{\lambda}_t}$ is the lambdas' standard deviation (Cochrane 2005, 245–252).

Petersen (2009) analyzes how different studies in respected finance journals estimate the standard errors and coefficients in a panel data set. Accordingly, many studies do not adjust standard errors to account for possibly cross-correlated residuals. Cross-sectional correlation causes standard errors to be understated and therefore overestimates the significance. Standard errors may require corrections

to be consistent in the presence of cross-sectional correlation. The Fama–MacBeth procedure is the most used method among studies adjusting standard errors. A fixed-effects regression approach uses dummy variables for each cluster to correct standard errors. The alternative method corrects standard errors for correlations within a cluster, and these clusters include firm, time, or both. Another method adjusts standard errors of OLS regressions with either White (1980) or Newey and West (1987) procedure (Petersen 2009). The Newey–West estimator adjusts for the heteroscedasticity and autocorrelation of the White procedure (Newey and West 1987).

According to Petersen (2009), clustered standard errors are robust to heteroscedasticity, and therefore the White standard errors may be more appropriate. According to Petersen (2009), many studies use incorrect adjustment methods. The most common panel data issues in finance applications include time-series dependence or firm effect and cross-sectional dependence or time effect. The residuals of a given firm correlate across time in the firm effect. Additionally, the residuals of a given time may correlate across different firms in the time effect. The level of standard errors should indicate any possible effects (Petersen 2009). The Fama–MacBeth procedure corrects for the time effect. However, the FMB may fail to address the firm effect in a data set. This issue causes standard errors to be biased downwards. High autocorrelation among factors may also result in biased standard errors (Petersen 2009). Suggested adjustments include the correction of Shanken (1992), for example. It corrects for issues when the beta estimators of Equation (16) come from the same sample as the generated independent variables. However, Petersen (2009) claims the FMB standard errors to be biased in most cases. Also, fixed effects regression only corrects for fixed effect biases (Petersen 2009).

Both Cochrane (2005, 245–252) and Petersen (2009) suggest clustering with multiple dimensions. This double clustering may be the most appropriate approach with sufficient data. The standard errors should be unbiased in the presence of both permanent and temporary firm effects (Petersen 2009). Brennan et al. (1998) suggest a variation of the Fama–MacBeth procedure, which deducts the factor loadings from the risk-adjusted returns before the second step. This procedure should correct for the errors-in-variables issue (Brennan et al. 1998). Another method to test regression models includes, for example, two-pass regression without the rolling procedure, generalized method of moments (GMM), and maximum likelihood (ML). The standard two-pass regression does not adjust for the cross-sectional correlation, unlike the FMB procedure. The GMM is essentially a modern two-pass regression considering possible conditional heteroscedasticity,

serial correlation, and non-normal distribution. The method is an appropriate alternative to Fama–MacBeth procedure. However, the FMB may be easier to implement even if the standard errors require adjustments. Maximum likelihood is a particular case of the GMM (Cochrane 2005, 235–278).

2.4 Market sentiment

As demonstrated in Equation (7), the risk factors, which proxy for the marginal utility or aggregate expected future cash flows in a more practical sense, drive asset returns (Subrahmanyam 2010). The changes in risk premiums can explain the majority of aggregate asset returns in the long run but may fail to do so in the short-term. Market sentiment, or investor attention, refers to the overall attitude towards the market's anticipated price development (Barberis et al. 1998). Accordingly, the behavior of investors can be quantified and used to explain asset returns. Especially retail investor behavior should be quantifiable. Therefore, the market sentiment combines aspects of behavioral finance with asset pricing. Different approaches to quantify investor attention include measures from the financial markets, surveys, text mining or news analytics, non-economic indicators, and internet search behavior (see, e.g., Barberis et al. 1998; Baker and Wurgler 2006; 2007). Market sentiment is an increasingly popular subject in academic studies, but the classic finance theory does not recognize it (Baker and Wurgler 2006).

Factors based on financial market-related measures include both fundamental and technical components. Also, market sentiment, in general, is often a contrarian indicator. De Bondt and Thaler (1985) suggest long-term past returns as a reversal strategy. Accordingly, a portfolio of stocks with over three cumulative years of poor performance outperforms the opposite portfolio going forward. Similar evidence from shorter periods also exists. Lo and MacKinlay (1990b) study a cross-sectional contrarian strategy, where they buy recently underperformed stocks and sells outperformed stocks. Possible explanations for this anomaly include the cross-autocorrelations across stocks and price overcorrection (Lo and MacKinlay 1990b). However, Fama and French (1996) argue that the FF3 model's factors explain the long-term reversal anomaly.

According to an alternative explanation for Fama and French (1993) factors, the market may misprice stocks with neglected characteristics. This mispricing provides long-term excess returns when prices adjust to fair value. You can attribute the factor performance to investor behavior. This alternative explanation

for Fama and French (1993) factors contradicts the EMH. Both Fama and French (1992) and Lakonishok, Shleifer and Vishny (1994) believe stock returns positively correlate to earnings-to-price and cash flow-to-price ratios. However, Lakonishok et al. (1994) argue the irrational earnings growth expectations towards so-called glamour stocks with low book-to-market characteristics explain the value premium. According to Barberis et al. (1998), investors either under or overreact to new information based on psychological evidence. Investors underestimate significant new information, such as earnings announcements. On the other hand, investors appear to overreact to a series of similar, either good or bad, news. New information reflects on prices more slowly than the EMH suggests. Consistent news patterns may drive asset prices to either overly high or low levels, resulting in a mean reversion (Barberis et al. 1998).

Unlike the contrarian approach, momentum strategies suggest trend following. The persistence of the momentum effect does not have a sensible explanation. Therefore, it is one of the biggest arguments against the efficient market hypothesis. A possible explanation for the momentum effect includes behavioral biases. However, it is not evident why the momentum works (see, e.g., Daniel, Hirshleifer and Subrahmanyam 1998; Jegadeesh and Titman 2001). The momentum effect usually refers to the cross-sectional factor described in Equation (10). However, you can also apply it to a time series strategy. Time-series momentum categorizes the securities based on their absolute performance, such as stocks with either positive or negative returns, instead of their relative performance. This strategy implies market timing because the number of companies in a portfolio depends on the market performance. The strategy offers consistent outperformance though it does not fit in the traditional asset pricing framework (Moskowitz, Ooi and Pedersen 2012).

Baker and Wurgler (2006; 2007) study proxies for market sentiment and suggest financial markets-based measures. These include trading and initial public offering (IPO) volumes, closed-end fund discount, first-day returns on IPOs, and implied volatility. For example, high trading volume indicates investors being optimistic, resulting in lower subsequent returns. Because of limited short-selling opportunities, retail investors impact trading volume more when they are net buyers. Implied market volatility measured with CBOE Volatility Index (VIX) has an opposite relation. Low VIX index levels proxy for the positive sentiment, which precedes a decline in asset prices (see Baker and Wurgler 2007).

Financial market-based measures also cover technical indicators, such as moving averages. These indicators are often subjective and, therefore, difficult to study (Lo, Mamaysky and Wang 2000). Technical analysis, in general, is not

widely recognized in academic literature (Fama 1970). However, Lo et al. (2000) objectively and systematically evaluate different technical analysis methods using pattern recognition techniques. They find evidence of several indicators providing at least some level of additional information. The advancement in algorithms may enhance technical analysis methods and help discover new anomalies (Lo et al. 2000).

Some sentiment proxies originate from surveys, such as consumer sentiment and consumer confidence indices. Brown and Cliff (2005) suggest survey-based sentiment to affect pricing. The results appear to be robust in a multi-factor asset pricing model. Accordingly, sentiment should be a contrarian indicator where high sentiment suggests inferior expected future returns and vice versa. You can categorize market sentiment as bullish, bearish, or neutral. However, no theoretically sound way to construct the sentiment index exists (Brown and Cliff 2005). According to Brown and Cliff (2005), the survey sentiment predicts the returns best over one to three years. The survey information can be lagged, often by weeks or months. In contrast, other sentiment measures become available at a much faster rate. Surveys can be subjective, and people may not behave as they would indicate (Baker and Wurgler 2007).

Text mining or news analytics extracts information from different textual data sources to form a proxy for a sentiment. The sources vary from traditional media platforms to social media. Antweiler and Frank (2004) study internet stock message boards and categorize the content of the messages. They found a relationship between the sentiment based on chat room activity and trading volume. There also appears to be a connection between the activity and market volatility. However, they do not find a significant relationship between the activity and returns. Tetlock (2007) measure the sentiment by analyzing Wall Street Journal (WSJ) articles' words. Accordingly, words used in the media and associated with pessimism impact trading volume and stock prices. The high media pessimism tends to forecast price reversion and increase in trading volume. Also, low market prices raise pessimism in the media. Bollen, Mao and Zeng (2011) argue that Twitter data can capture investors' moods. They find the sentiment based on Twitter posts or tweets to be a reliable predictor of future price movement and volatility.

Seemingly unrelated events may influence our mood, behavior, and risk preferences. Non-economic factors, such as weather conditions, may capture investor attention. Hirshleifer and Shumway (2003) believe the festive mood resulted from sunny weather influences the market. According to their study, the morning sunshine positively correlates with stock returns, and the effect is significant. However, sunshine, rain, and snow do not have a similar impact. According to Cao and

Wei (2005), the temperature affects investors' mood, and stock returns negatively correlate with the temperature. Therefore, cold weather leads to higher returns, and the results appear to be robust when controlled with other factors. Other non-financial factors influencing investor behavior include, for example, changes in daylight savings time, lunar phases, and the length of the day and night (see, e.g., Cao and Wei 2005). You can explain the results with psychological effects (Hirshleifer and Shumway 2003).

An increasingly popular approach quantifies market sentiment with internet search behavior. Google Trends search volume data is a popular method to capture investor attention. The availability and the amount of Google Trends data make it an appealing source for sentiment proxies. Google accounts for most internet searches, especially in the United States. The Google data may capture less sophisticated retail investors' behavior (see, e.g., Da et al. 2011). Moat, Curme, Avakian, Kenett, Stanley and Preis (2013) study the views and edits of Wikipedia articles. They find the increase in page visits in financially related articles to anticipate a decrease in market prices.

Although the studies about market sentiment show promising results, the studies focus on sentiment as a trading strategy. Therefore, the result and the methodology used may not apply to asset pricing. Chordia et al. (2017) conduct a comprehensive meta-analysis on trading strategies and find serious issues related to their methods. Accordingly, many studies fail to test the hypothesis correctly, which results in many false-positive results. Few trading strategies outperform the market (Chordia et al. 2017).

2.5 Google Trends search volume index

Internet search behavior is a measure for investor attention. Retail investors use the internet as an information source before making a transaction (Da et al. 2011). Preis et al. (2012) study Google search usage and find internet search behavior more future-orientated in countries with high gross domestic products. Therefore, Google search data can be a useful predictor of future events. Google Trends data may reflect subsequent stages in the decision-making process of investors. An increase in Google search volumes for keywords relating to financial markets may work as a warning signal for stock market falls (Preis et al. 2013). According to Preis et al. (2013), you can then use these warning signs to construct profitable trading strategies.

Google Trend data is available as a search volume index (SVI). It shows how often people search for a particular term on Google relative to the total search volume. Search query information from web searches, as well as from other Google's services, can be observed. Google Trends also allows filtering the results based on, for example, geographic region and period. The normalized data is a proportion of all the searches made at time and location. This normalization enables to compare the interest over topics in different geographic areas and adjusts to the overall increase in search volume. The SVI values range from zero to 100, where 100 is the maximum interest for a search term in a selected place and time (Stephens-Davidowitz and Varian 2014).

Choi and Varian (2012) demonstrate how to forecast different near-term economic indicators with Google Trends before the release of official figures. This method of forecasting is also known as nowcasting. The tested indicators include automobile sales, unemployment claims, travel destination plans, and consumer confidence. Including Google Trends data in an autoregressive model is believed to increase its prediction capabilities (Choi and Varian 2012). Several studies examine web search data in various fields. A notable study by Ginsberg, Mohebbi, Patel, Brammer, Smolinski and Brilliant (2009) demonstrates how Google Trends data can detect influenza epidemics. People tend to search for information about their symptoms. You can reliably track the spread of an epidemic with the data. This method does not suffer from a usual reporting lag (Ginsberg et al. 2009).

Barber and Odean (2008) believe retail investors to be net buyers of attention-driven stocks. Retail investors may struggle to choose which stock to buy and tend to lean towards the ones capturing their attention. This finding applies especially to investors trading with large discount brokerage services. Retail investors must sell stocks they already own because of limited short-selling opportunities (Barber and Odean 2008). Da et al. (2011) test this hypothesis with SVI data as a proxy for attention. When an investor searches for information about a particular stock, it is a definite measure of attention. Therefore, the SVI can measure investor attention directly, unlike other market sentiment measures.

Da et al. (2011) study SVIs of different stock ticker symbols and find a rising SVI value to anticipate a subsequent drop in prices. The attention measured by the SVI data can also explain the long-term underperformance of initial public offerings (Da et al. 2011). Preis, Reith and Stanley (2010) find a link between SVI volume and trading volume. Accordingly, weekly transaction volumes of companies correlate with the SVI volumes of the corresponding company names. The more people search for companies, the more they also trade them. Therefore, the search volume reflects the attractiveness of a stock. However, the high trading volume

may also increase both the attention and the SVI volumes (Preis et al. 2010). Dimpfl and Jank (2016) find a similar relationship between Google search volume and volatility. The increasing volume raises investor attention, and great attention also leads to higher volatility.

Preis et al. (2013) study 98 different search terms of varying financial relevance to determine if the SVI volume of such keywords anticipates moves in the stock market. Search terms related to financial markets anticipate future trends and can be used to construct profitable trading strategies. The best performing keyword in the study of Preis et al. (2013) is *debt*. Preis et al. (2013) test a trading strategy based on the weekly search occurrence of keywords. If the search volume is lower than the three-week average, they buy the index. Consequently, if the search volume is higher than the average three-week search volume, they take a short position on the index. The holding period is one week for both long and short positions.

Trading based on the keyword *debt* generates a cumulative return of 326% between January 2004 and February 2011 compared to the 16% of the buy-and-hold strategy during the same period. The result is also over two standard deviations above a trading strategy based on simulated random "buy" or "sell" decisions (Preis et al. 2013). Preis et al. (2013) also analyze the characteristics of successful keywords. They find a high positive correlation between the search term's financial relevance and its likelihood to forecast subsequent stock market moves. They measure this relevance based on the appearances in Financial Times concerning the search term's popularity overall. The relevance effect is also statistically significant (Preis et al. 2013).

According to Curme, Preis, Stanley and Moat (2014), the predictive value of the keywords Preis et al. (2013) suggests can diminish over time. Curme et al. (2014) further develop the study of Preis et al. (2013) using semantic categories instead of individual keywords. They create topics with computational linguistics techniques to retrieve the SVI data. The increase in interest in politics and business-related topics may forecast declines in the stock market. Interest in these topics should reflect concerns about the economy. However, the relation between the Google Tends data, and the stock market appears to weaken. Relevant information may still be found in the trends of less-obvious keywords and categories (Curme et al. 2014).

Da et al. (2015) form a Google Trends-based fear sentiment from the search terms related to commons concerns, such as recession, unemployment, and bankruptcy. They use dictionaries to categorize search terms as either positive or negative objectively. This form of market sentiment predicts price reversals,

volatility spikes, and fund flows. Gold price and recession-related words appear to be the most significant (Da et al. 2015). Da et al. (2015) use principal component analysis to form the sentiment.

2.6 Principal component analysis in sentiment extraction

Principal component analysis (PCA) is a multivariate analysis technique for finding patterns in large data sets. PCA can reduce the dimensions of the data and extract the most relevant information. This technique obtains the components from a data matrix based on their eigenvalues. PCA can reduce the variance dimensions and reduce the data's noise (Abdi and Williams 2010). The first principal component is the direction of maximum variance. The second principal component is the direction of maximum variance perpendicular to the first principal component's direction and so on. The first components usually contain the most relevant information (Abdi and Williams 2010).

The principal component analysis refers to an orthogonal decomposition. A standard way of presenting a matrix in linear algebra is to use eigenvectors and eigenvalues. Eigenvectors or characteristic vectors resemble the principal components, and eigenvalues refer to the variance they explain (Abdi and Williams 2010). A standard method is to determine the eigenvector of matrix $\mathbf{A}$ as a vector $\mathbf{u}$. The vector fulfills the following eigenvector equation

$$\mathbf{Au} = \lambda\mathbf{u}, \tag{19}$$

where λ is a scalar or eigenvalue associated with the eigenvector. You can rewrite Equation (19) as

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{u} = \mathbf{0},$$

where $\mathbf{I}$ is the identity matrix, and $\mathbf{0}$ is the zero factor. In PCA, you can obtain the loadings as

$$Loadings = Eigenvecors \times \sqrt{Eigenvalues}. \tag{20}$$

Loading refers to the correlation between a component and a variable estimate (Abdi and Williams 2010).

Baker and Wurgler (2006) study how the market sentiment explains the cross-section of asset returns, as described in Section 2.4. They suggest several financial market-related measures for proxies. They form a sentiment index from multiple measures using PCA instead of individual proxies for the sentiment. This method

should isolate the standard components of different criteria. Baker and Wurgler (2006) use six measures and their lags and define sentiment index as the first principal component of this data set. The first principal component explains about half of the sample variance. Then, they form an alternative sentiment index from the residuals, which also explains about half of the variance. When the sentiment is high, stocks earn lower subsequent returns (Baker and Wurgler 2006).

According to Baker and Wurgler (2006), you cannot capture the complex financial markets and investor behavior with any single sentiment proxy. Later, they expand this study to test the relation between firm characteristics and sentiment. The sentiment most likely affects new and unprofitable growth firms with high profitability potential. On the contrary, it is less likely to affect value firms with a long history of earnings, tangible assets, and stable dividends (Baker and Wurgler 2007).

Tetlock (2007) uses principal component analysis to form a sentiment based on the WSJ columns' pessimism. PCA should extract the most important semantic information. The first principal component appears to capture media pessimism quite well. As described in Section 2.4, the high media pessimism tends to forecast price reversion and increasing trading volume. Constructing a trading strategy based on media pessimism would yield insignificant abnormal returns and generate meaningful real-world costs due to a high portfolio turnover. Limits to sentiment arbitrage may also prevent the market from responding efficiently to sentiment information (Tetlock 2007).

Zhang (2009) extracts the FF3 factors from Fama–French portfolios formed by size and book-to-market factors using PCA. They find the extracted factors to explain the stock returns better than the original factors. Therefore, PCA may be a valid method in asset pricing applications. However, the extraction method only works with the portfolio returns and fails with individual stocks (Zhang 2009).

Da et al. (2015) combine PCA and Google Trends search volume data. As described in Section 2.4, they create an aggregate fear sentiment index from negatively associated keywords. Da et al. (2015) find this form of market sentiment to predict price reversals, volatility spikes, and fund flows (Da et al. 2015). Big data offers possibilities to understand the cross-section of asset returns better. However, it also requires appropriate methods to extract relevant information. Principal component analysis can be useful for this purpose.

3 DATA AND METHODOLOGY

3.1 Description of the NYSE data

I acquire NYSE data from Thomson Reuters Datastream using the constituent list *FUSNYSE*. The data includes both return data and some stock-specific characteristics. I calculate percentage returns from the respective total return index (*RI*), which expresses a theoretical growth in the value and includes reinvested dividends, for example. The characteristics include market value (*MV*), price-to-book value ratio (*PTBV*), the price-to-earnings ratio (*PE*), price-to-cash flow ratio (*PC*), and dividend yield (*DY*). I filter firms based on SIC code 1 (*WC07021*). SIC code stands for standard industry classification code and SIC code 1 represents a business segment providing the most revenue for a company (Thomson Reuters 2015, 641). I identify stocks based on their Datastream codes.

The monthly data ranges from 2004 to 2017, and I select this range due to the limitations of Google Trends data. The list *FUSNYSE* includes 2,876 companies. First, I filter out companies without a SIC code. Second, I remove all companies in a SIC code range of 6000–6799. These include finance, insurance, and real estate-related companies. Third, I remove companies if they do not have total returns data or lack data for one or more characteristics. However, I do not remove companies even if they lack data for returns or characteristics data for occasional months to have a decent sample size. I differentiate the stocks with their Datastream codes. The data set includes 2,120 firms after the first step, 1,548 after the second step, and the cleaned data set contain 1,290 firms.

I construct cross-sectional factors from the NYSE data and call them NYSE factors. I sort the firms into ten deciles based on these characteristics. I then calculate the factors as an average equal-weighted return of the companies in the smallest three deciles less respective returns of the companies in the largest three deciles. The dividend yield factor differs from other factors. I define it as the average equal-weighted return of dividend-paying companies minus the returns of companies not paying dividends. I calculate the factors for each month. I adopt this approach from Baker and Wurgler (2006), although they used slightly different characteristics. The characteristics consist of commonly used figures or ratios to categorize stocks. I chose the characteristics for self-constructed cross-sectional factors based on the availability of the Datastream data.

The NYSE data suffers from issues that may affect the results. I must filter out many stocks due to missing data. Also, the data may suffer from survivorship since

Datastream does not specify delisted stocks within the period. I must neglect this bias because I find it difficult to avoid. Some of the characteristics may also suffer from look-ahead bias. For instance, Datastream may use forward-looking earnings to calculate price-to-earnings ratios (Thomson Reuters 2015, 40–41, 451–452).

3.2 Fama–French portfolios and factors

I obtain Fama–French portfolios and factors from the Data Library of French (2018). I choose this data because of its popularity in asset pricing testing. I use average value-weighted excess returns of 25 portfolios formed on size and book-to-market as my left-hand variables in regressions. Asset pricing studies commonly use these portfolios in empirical tests (see, e.g., Fama and French 1993). Right-hand variables contain the risk of the five-factor model of Fama and French (2015). These factors include excess market return ($R_m - R_f$), small minus big (SMB), high minus low (HML), robust minus weak (RMW), and conservative minus aggressive (CMA). I use the factors in Equation (11).

Each time series consist of 168 monthly observations from January 2004 to December 2017. I select this range due to the limitations of Google Trends data, which is available from 2004 onward. The Data Library data originates from the Center for Research in Security Prices (CRSP) database except for the one-month Treasury bill rate (T-bill). It is used as a risk-free rate and comes from Ibbotson Associates (French 2018). The Fama–French portfolios include stocks, NYSE, AMEX, and NASDAQ stocks with sufficient data.

The ($R_m - R_f$) is the value-weighted return of CRSP firms minus the one-month T-bill rate. Also, the Fama–French risk factor construction uses equal-weighted returns of different value-weighted portfolios. The SMB uses market equity, and HML depends on book-to-market ratios. RMW refers to profitability, and CMA relates to investing activities. French (2018) defines the SMB factor as

$$SMB = \frac{1}{3}(SMB_{BM} + SMB_{OP} + SMB_{INV}),$$

where SMB is the average return difference of the nine small and nine big stock portfolios. Portfolios SMB_{BM}, SMB_{OP}, and SMB_{INV} each consist of an average return difference of three small and big portfolios. The HML is

$$\begin{aligned} HML = &\frac{1}{2}(Small\ Value + Big\ Value) \\ &- \frac{1}{2}(Small\ Growth + Big\ Growth), \end{aligned}$$

where *HML* is the equal-weighted return difference of value and growth portfolios. The *RMW* factor is

$$RMW = \frac{1}{2}(Small\ Robust + Big\ Robust) - \frac{1}{2}(Small\ Weak + Big\ Weak),$$

where *RMW* bases on the returns of robust and weak portfolios measured by their profitability. The *CMA* factor is

$$CMA = \frac{1}{2}(Small\ Conservative + Big\ Conservative) - \frac{1}{2}(Small\ Aggressive + Big\ Aggressive),$$

where *CMA* is the average return spread between conservative investment aggressive investment portfolios (French 2018). Fama and French (2015) categorize portfolios into conservative and aggressive based on total assets change.

3.3 Google Trends data

I select the Google Trends data based on 98 different search terms Preis et al. (2013) propose. Preis et al. (2013) choose terms related to stock markets using the Google Sets service, a tool to identify semantically related keywords. According to Preis et al. (2013), the terms include some intentional financial bias. I do not pick my own words to avoid selection bias and to increase the study's objectivity. This setting also enables to test the research of Preis et al. (2013) with a more comprehensive methodology. In this study, I limit the data source of Google Trends to web searches only.

I obtain the data with a statistical program R and provide the script for retrieving the Google Trends data in Appendix 1. I use R package *gtrendsR* by Massicotte and Eddelbuettel (2018) to retrieve Google Trends queries individually for each keyword. I then merge the data. This way, one search term's relative occurrence does not affect the values of other search terms. As described in Section 3.3, the SVI values range between zero and 100, where the value 100 represents the highest relative occurrence. Every search query's time series contains 168 monthly observations ranging from January 2004 to December 2017, correspondingly to the financial time series used in this study. Google Trends only provides monthly data for such a lengthy period (Stephens-Davidowitz and Varian 2014). This data

frequency should not be an issue since empirical asset pricing often uses monthly frequency. The data also contains less statistical noise.

Geographically, I limit both Google data and financial data to the United States. As described in Section 3.3, Preis et al. (2012) study Google search among households. They find internet search behavior to be future-orientated in countries with a high gross domestic product, such as the United States. Also, the English language makes testing different search terms easier. Many people in the United States likely use English when searching for information online. Also, the U.S. stock market is the largest globally. People have broad access to the internet, and they likely invest in their domestic market.

Google Trends data depends on a sample representative to all Google searches since analyzing all the search data would not be plausible. This random sample updates daily, and therefore, the SVI values may also vary daily. According to Stephens-Davidowitz and Varian (2014), sampling should give reasonably precise estimates and does not necessarily require averaging over multiple samples. However, Google Trends rounds the search volume in each instance to the nearest integer. Averaging over various instances should then result in more precise estimates (Stephens-Davidowitz and Varian 2014). I averaged the SVI values over 15 different realizations to increase the accuracy.

3.4 Constructing the sentiment indices

I use principal component analysis to construct four different sentiment indices based on Google Trends data. The first sentiment index contains all the search queries used in the study of Preis et al. (2013). The second index includes 15 best-performing keywords. The third index comprises 15 queries with the highest relative keyword occurrence, according to Preis et al. (2013). The fourth sentiment index only contains the best performing keyword *debt*, according to the same study. I arrange the search queries in descending order based on their performance. The order may affect the direction of the sentiment index when using principal component analysis. Using the best search queries of Preis et al. (2013), I am giving the Google Trends data an ideal chance to succeed when tested in an asset pricing context.

First, I normalize the data by taking a logarithmic difference. I omit search query *rare earths* from the data set since its search volume of zero does not allow the calculation of logarithmic differences. I lose the first observations of the search

volume data when I calculate the difference. I include lagged sentiment indices for the four different sentiments with a lag of one. When delaying variables, I lose the last observations of sentiment indices. The second step transforms the data into a covariance matrix. It describes the variance of the data and the covariance among different variables. Finally, I identify the eigenvectors and eigenvalues of this covariance matrix. The loadings of the principal components can be expressed as eigenvectors times the square root of the eigenvalues. I describe the loadings also in Equation (20).

Finally, I take a scalar product of the original data matrix and the first principal component's loadings. The first principal component is the direction of maximum variance. The second principal component would be the direction of maximum variance perpendicular to the first principal component's direction. The number of components corresponds to the dimensions of data. However, I only use the first component that explains the majority of the variance. The purpose of using principal component analysis is to extract the essential information and remove noise. However, some information may be lost when using only one of the components. Thus, the four indices constructed using this methodology should be adequate proxies for the sentiment.

3.5 Regression methods and hypothesis testing

I use various regression methods to answer the research questions and reach the aim. The methods include Fama–MacBeth regressions, regressions with clustering, long-short regressions, and risk-adjusted regressions. Fama–MacBeth regressions relate to the second research question of whether the Google Trends-based sentiment explains the cross-section of asset returns. The FMB is the thesis's primary regression method, and I perform it on both NYSE stocks and Fama–French portfolios. The rest of the methods relate more to the first research question of whether the synthetic sentiment indices constructed in this thesis is related to aggregate stock market movements. I perform these methods only on NYSE stocks since they require stock-specific characteristics described in Section 3.1. I use clustering to address some standard error correction issues associated with the Fama–MacBeth methodology (see Petersen 2009).

The two-step methodology of Fama and MacBeth (1973), which I describe more profoundly in Section 2.3, estimates the risk factors' betas and risk premiums. First, I structure the data to a panel in a long format. I then add unique time and

asset indices for data filtering and rearrangement purposes. The data processing steps described in Section 3.1 result in an unequal number of observations per firm, which causes the panel of NYSE stocks to be unbalanced. The panel of Fama–French portfolios, on the other hand, is balanced. I also omit all panel rows with missing data for at least one of the columns.

In the first step, as described in Equation (12), I arrange the panel based on assets and then based on time. I regress each asset's excess return against one of the sentiment indices and Fama–French five factors. This step determines each asset's rolling betas for risk factors, and I demonstrate the beta estimates in Equation (13). I use a time window of 60 months for rolling regressions, which aims to generate robust regressions without compromising too much data. Brennan et al. (1998) also use the same observation interval. NYSE data limitations require making a deliberate compromise between the amount and the quality of the regressions. Therefore, an asset must have at least 45 out of 60 valid observations for regressions.

The first Fama–MacBeth regression step with sentiment and FF5 factors is,

$$\begin{aligned} R_{i,t} - R_{f,t} = \alpha_i &+ \beta_{i,Sentiment} Sentiment_t + \beta_{i,(R_m - R_f)}(R_m - R_f)_t \\ &+ \beta_{i,SMB} SMB_t + \beta_{i,HML} HML_t + \beta_{i,RMW} RMW_t \\ &+ \beta_{i,CMA} CMA_t + \epsilon_{i,t}, \end{aligned} \tag{21}$$

where $Sentiment_t$ is one of the four sentiment factors as explained in Section 3.4. I also use lagged sentiments where I replace $Sentiment_t$ with $Sentiment_{t-1}$. In the second step, I merge the rolling betas horizontally to the panel. I rearrange the data primarily by time and secondarily by assets. I also filter out data points with no estimates. I regress asset's excess returns for a fixed period against the estimated betas to determine the risk premiums for factors. I present the essential second step in Equation (15).

The second step with sentiment and FF5 factors is,

$$\begin{aligned} R_{i,t} - R_{f,t} = \alpha_i &+ \lambda_{t,Sentiment} \hat{\beta}_{i,Sentiment} \lambda_{t,(R_m - R_f)} \hat{\beta}_{i,(R_m - R_f)} \\ &+ \lambda_{t,SMB} \hat{\beta}_{i,SMB} + \lambda_{t,HML} \hat{\beta}_{i,HML} + \lambda_{t,RMW} \hat{\beta}_{i,RMW} \\ &+ \lambda_{t,CMA} \hat{\beta}_{i,CMA} + \epsilon_{i,t}, \end{aligned} \tag{22}$$

where beta estimators come from Equation (21). Risk premium estimators of factors represent the mean of lambda coefficients. I demonstrate the estimator in Equation (17). The lambda estimates represent factor risk premiums for a fixed period. I average monthly risk premiums of a factor to derive the risk premium estimator for factors. I compute the t-statistics manually from lambda coefficients

as described in Equation (18). I perform these steps for each of the four sentiment indices individually.

I test if the lambda coefficients are significantly different from zero. The null hypothesis is

$$\begin{aligned} H_0 &: \lambda_k = 0 \\ H_1 &: \lambda_k \neq 0, \end{aligned} \tag{23}$$

where λ_k is the lambda coefficient of a factor k, lambdas are zero in the null hypothesis and non-zero in the alternative hypothesis. I am mostly interested in the lambdas of each of the sentiments. Those lambda coefficients' significance answers the research question of whether the Google Trends-based sentiment explains the cross-section of asset returns. The t-statistics are also more relevant for this thesis than the coefficients. Suppose I reject the null hypothesis for sentiment factors. In that case, the data suggests Google Trends-based sentiment to explain the cross-section of asset returns.

Panel data may cause issues covered in Section 2.3. The residuals may be correlated across assets or time and be biased when using the Fama–MacBeth procedure. Possible alternatives include, for example, fixed effects and clustered regressions. Appropriate clustering may be the safest option for unbiased standard errors when the firm effect is temporary (Petersen 2009). Because of my panel's possible issues, I find clustering the best alternative to Fama–MacBeth methodology. Besides, I use White standard errors, which correct for heteroscedasticity (White 1980). I also test for a possible firm, time, or firm and time effect in the panel. I use each of the sentiment indices and NYSE factors covered in Section 3.1 as dependent variables.

I also use so-called long-short regressions inspired by Baker and Wurgler (2006). They test their sentiment indices against their self-made portfolios. Although slightly more simplified, I do the same with Google Trends-based sentiment indices and NYSE factors. I use the NYSE factors, which are essentially long-short portfolios, as dependent variables. I also include excess stock returns as one of the dependent variables for comparison. The independent variables include sentiment indices, which I use with and without the FF5 factors. I use White standard errors and appropriate clustering to mitigate possible issues in the data.

I also use a risk-adjusted regression method inspired by Brennan et al. (1998). First, I perform the first step of the Fama–MacBeth procedure. Instead of doing the second step like in Equation (22), I multiply the FF5 factors with their corresponding beta estimators and then deduct them from the excess return. I specify

this variation as

$$R^*_{i,t} \equiv R_{i,t} - R_{f,t} - \sum_{k=1}^{n} \hat{\beta}_{i,k} F_{k,t}, \tag{24}$$

where $F_{k,t}$ is a Fama–French factor k in month t. The risk-adjusted return $R^*_{i,t}$ defined in Equation (24) is then used as a left-hand variable

$$R^*_{i,t} = \alpha_i + \hat{\beta}_{i,t,Sentiment} + \sum_{k=1}^{n} Z_{k,i,t} + \epsilon_{i,t}, \tag{25}$$

where $Z_{k,i,t}$ is the value of characteristic k for asset i in month t. The characteristics used as right-hand variables in Equation (25) include the five NYSE characteristics explained in Section 3.1. I also use the beta estimators of sentiment indices as independent variables. I select beta estimators because the sentiment indices, unlike the characteristics, are not asset-specific.

4 RESULTS AND DISCUSSION

4.1 Sample descriptive statistics

Figure 1 shows both monthly and cumulative market returns in percentages, and these returns represent the total U.S. stock market. I calculate market return by adding the risk-free rate to the $R_m - R_f$ factor.

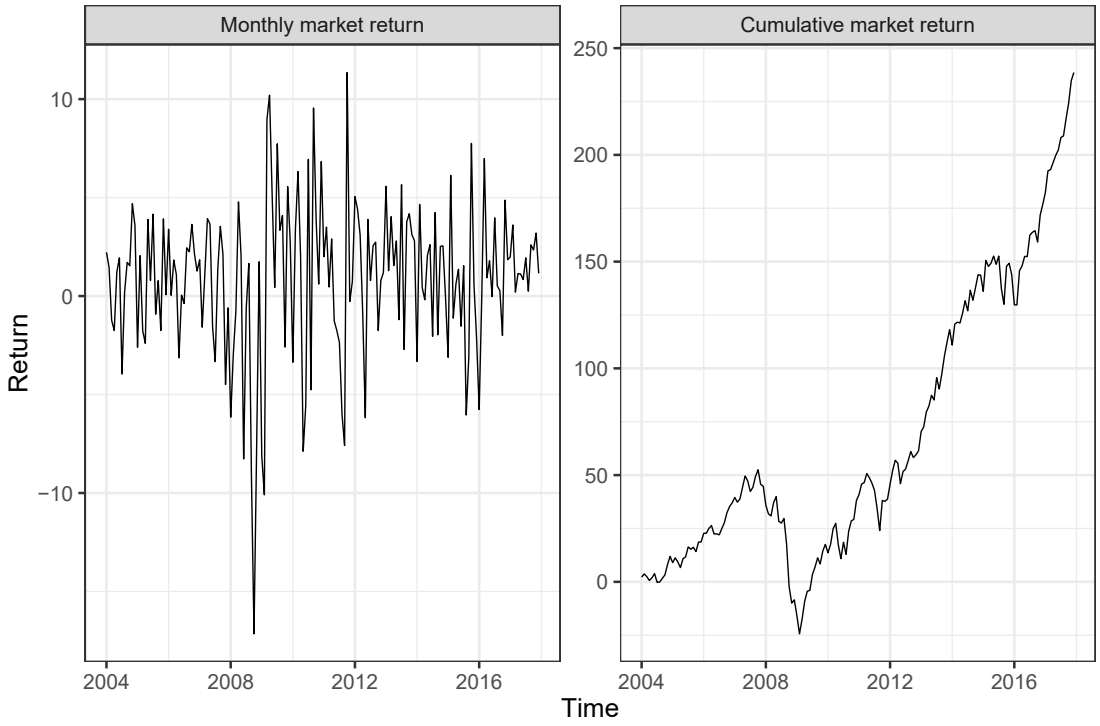

Figure 1: Market return

During 14-year the observation period, the stock market has a cumulative yield of 238.6%. The return equals an average annual return of 9.1 percent. Returns are favorable, especially on the second half of the sample, despite events, such as the financial crisis of 2007–2008 and the European debt crisis in the 2010s. Also, the risk-free rate and volatility are low, as described in Table 1.

In Figure 2, I select examples of different search volume indices. Market return in Figure 1 and the search volume index examples in Figure 2 have some resemblance.

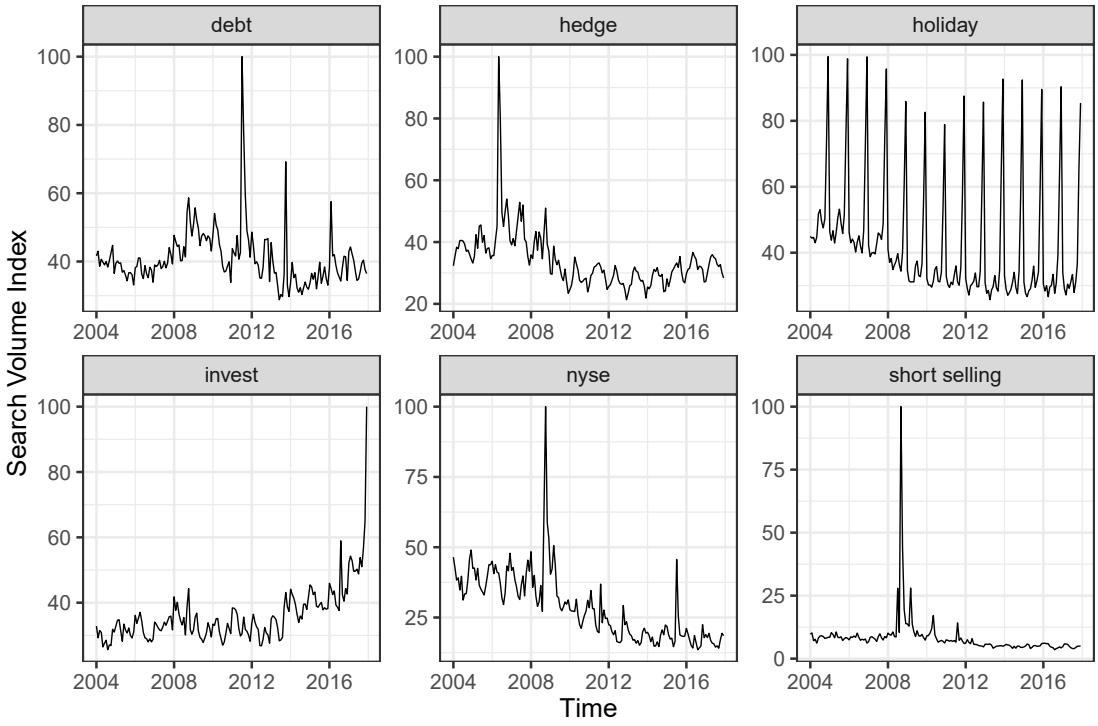

Figure 2: Search volume index examples

The search volume index of the term *debt* appears to peak around 2011 during the European debt crisis. Interestingly, term *hedge* peaks before the year 2007 and the financial crisis. The search volume of the keyword *holiday* is an example of a search term with clear seasonality. It experiences an increase in search volume every year around the holiday season. Search term *invest* ascends towards the end of the period. Interest in investing might increase when the stock market performs well and decrease during stock market crises.

On the contrary, terms *nyse* and *short selling* have a clear peek around the financial crisis. The search volume of *nyse* is almost the total opposite of the cumulative market return. The stock market appears to receive more attention during market crashes. During a financial crisis, there is also more market coverage in the news. Therefore, the increase in search volume may originate from people who do not follow the market regularly.

The following Figure 3 shows all the search queries Preis et al. (2013) propose and used in this study. As mentioned in Section 3.3, Preis et al. (2013) choose terms suggested by the Google Sets service with some bias towards the stock market. I sort the search terms by their median search volume in descending order. The vertical dashed line represents the mean of the sample. The lower and upper hinges of the boxplots correspond to the 25th and 75th percentiles.

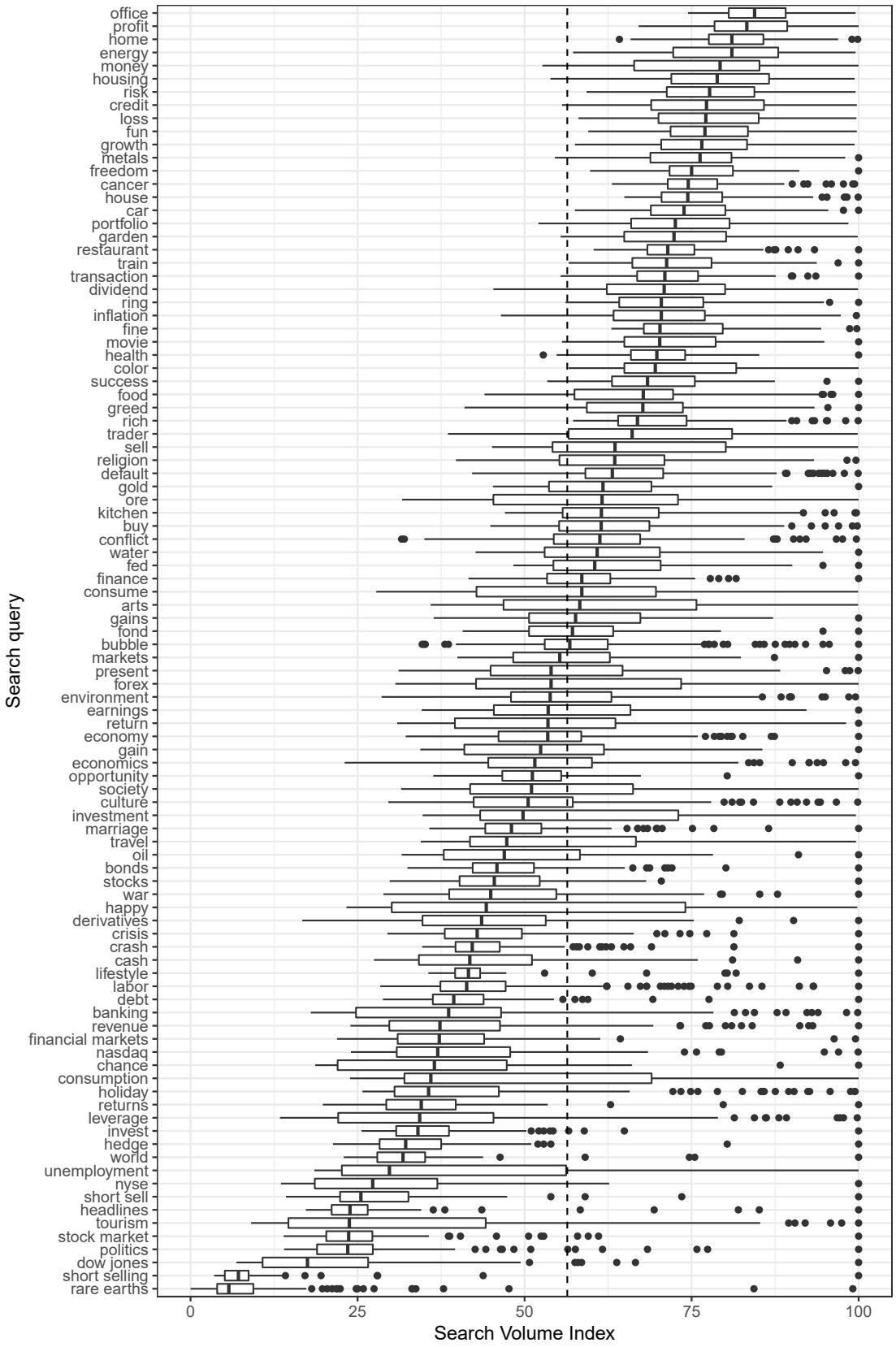

Figure 3: Search volume indices

Stock market-specific search terms seem to be less searched on average than the more general terms. All the search terms are included in the first sentiment index except *rare earths* because it does not always have a search volume. The second sentiment index consists of the 15 best-performing search queries according to Preis et al. (2013). These are *debt*, *color*, *stocks*, *restaurant*, *portfolio*, *inflation*, *housing*, *dow jones*, *revenue*, *economics*, *credit*, *markets*, *return*, *unemployment*, and *money*. Terms *color* and *restaurant* stand out because they are not related to the stock market.

The third sentiment index consists of the 15 most financially relevant search queries, according to Preis et al. (2013). These include *hedge*, *dividend*, *earnings*, *inflation*, *markets*, *bonds*, *debt*, *financial markets*, *gains*, *investment*, *growth*, *derivatives*, *crisis*, *unemployment* and, *banking*. The fourth sentiment index contains only the search term *debt*, which is the best performing search query according to Preis et al. (2013).

As explained in Section 3.4, I use principal component analysis to construct the sentiment indices. Figure 4 demonstrates how much variance different principal components explain. The more observations in a sample, the more the first component explains. The variance proportions of the components sum up to 100% for each of the sentiment indices.

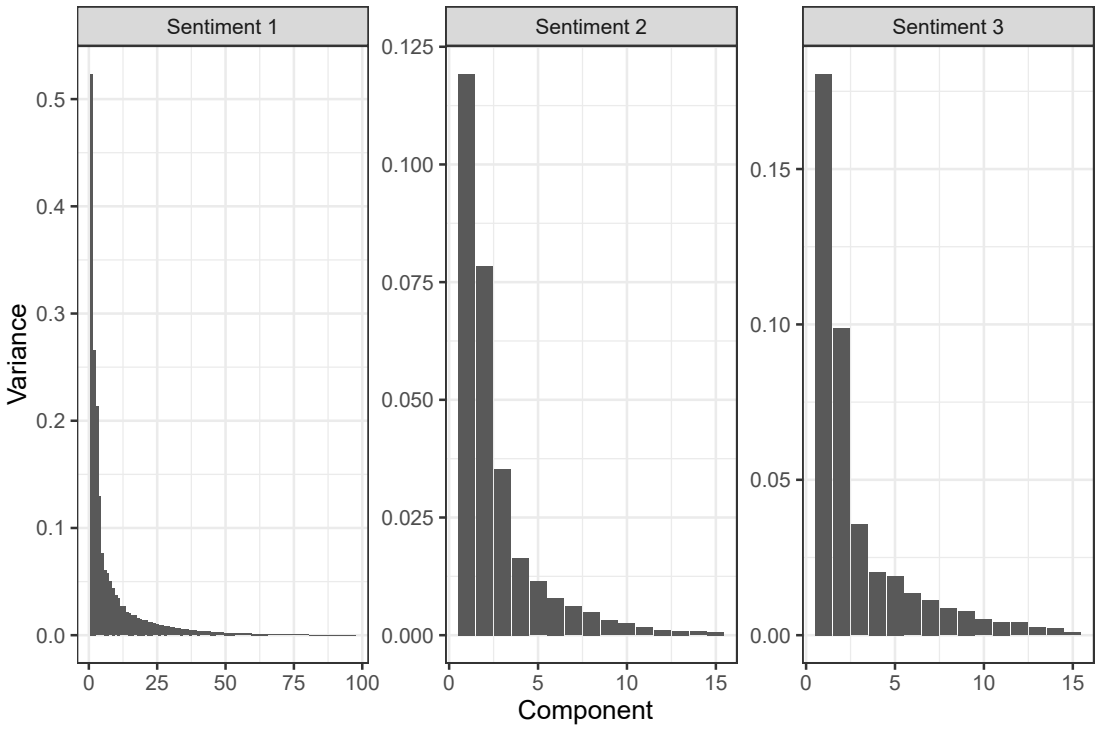

Figure 4: Variance proportions of principal components

The first principal component explains 52.3% of the first sentiment index's variance. For reference, the second component explains 26.6%, and the third component explains 21.3% of the variance. In the second sentiment index, the first component explains 11.9% of the variance. The first principal component explains 18.1% of the third sentiment index's variance. The fourth sentiment index is not included in Figure 4 since *debt* is the only search query of the sentiment index.

The four sentiment indices have some similar seasonality in Figure 5. I describe how to construct the sentiment indices in Section 3.4.

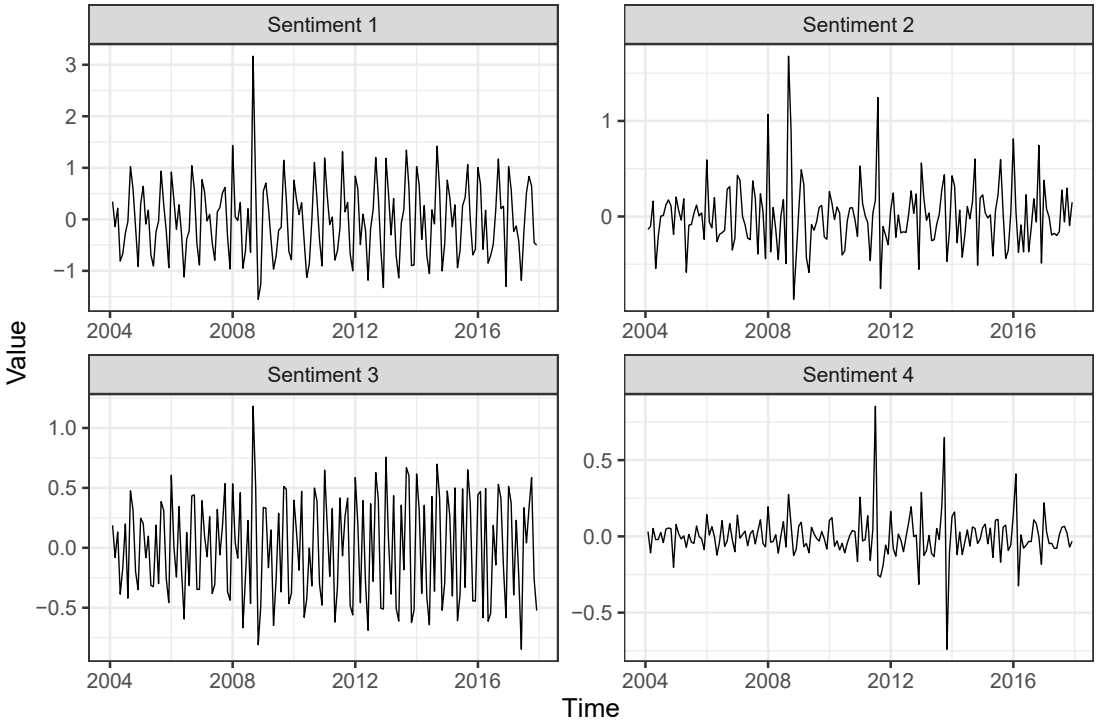

Figure 5: Sentiment indices

Sentiments 1 to 3 peek at the end of the financial crisis. The fourth sentiment index, based on search query *debt*, peeks around the European debt crisis. The more search queries there are in the sentiment index, the more variance there is. Figure 5 reflects a monthly change in the sentiment instead of cumulative development, so the trend is not apparent. The scale expresses values and not a percentage or logarithmic change. This scale explains why the first sentiment index has values below -1 at times. I include a correlation matrix of the sentiments in included in Table 2. Table 1 presents the descriptive statistics of the data apart from Fama–French 25 portfolios.

Table 1: Data descriptive statistics

This table presents the descriptive data statistics. The NYSE stock returns, characteristics, and factors are defined as described in Section 3.1. I define the Fama–French data in Section 3.2 and present the sentiment data in Section 3.4. I round the figures above 1,000 in NYSE stock returns and characteristics to the nearest integer. The market value (MV) represents USD millions.

NYSE stock returns and characteristics

Statistic	N	Mean	SD	Min	Pctl(25)	Pctl(75)	Max
RI	171,840	1.231	11.559	−87.764	−4.317	6.364	389.371
MV	172,407	12,375	31,325	0.390	837.120	9,272	513,362
PB	172,014	1.900	234.430	−31,854	1.350	3.390	3,241
PE	144,796	41.048	591.363	0.000	13.700	26.400	101,750
PC	171,880	7.672	272.030	−29,745	6.000	13.050	8,767
DY	172,403	1.924	3.003	0.000	0.000	2.740	142.860

NYSE factors

Statistic	N	Mean	SD	Min	Pctl(25)	Pctl(75)	Max
MV SMB	168	−0.283	3.013	−8.723	−1.763	1.524	12.635
PB Low−High	168	−1.879	2.300	−12.597	−2.744	−0.439	6.597
PE Low−High	168	−2.275	2.376	−14.634	−3.255	−0.949	5.526
PC Low−High	168	−1.528	2.660	−14.228	−2.862	0.116	5.839
DY $>0-=0$	168	−0.236	1.784	−6.085	−1.373	0.768	6.245

Fama–French five factors and a risk-free rate

Statistic	N	Mean	SD	Min	Pctl(25)	Pctl(75)	Max
$R_m - R_f$	168	0.711	3.977	−17.230	−1.282	3.120	11.350
SMB	168	0.126	2.372	−4.790	−1.415	1.680	6.930
HML	168	0.033	2.516	−11.100	−1.265	1.282	8.270
RMW	168	0.297	1.634	−4.030	−0.820	1.180	4.900
CMA	168	−0.021	1.404	−3.340	−1.045	0.865	3.670
RF	168	0.097	0.140	0.000	0.000	0.152	0.440

Sentiments and lagged sentiments

Statistic	N	Mean	SD	Min	Pctl(25)	Pctl(75)	Max
Sentiment 1	167	−0.010	0.726	−1.552	−0.595	0.486	3.166
Sentiment 2	167	0.002	0.346	−0.867	−0.198	0.171	1.678
Sentiment 3	167	−0.004	0.426	−0.847	−0.381	0.388	1.182
Sentiment 4	167	−0.001	0.146	−0.740	−0.068	0.056	0.853
Sentiment 1 (−1)	166	−0.007	0.727	−1.552	−0.600	0.494	3.166
Sentiment 2 (−1)	166	0.001	0.347	−0.867	−0.198	0.172	1.678
Sentiment 3 (−1)	166	−0.001	0.426	−0.847	−0.380	0.389	1.182
Sentiment 4 (−1)	166	−0.001	0.147	−0.740	−0.068	0.056	0.853

The N of Table 1 shows the number of observations for each of the time series. NYSE stock returns and characteristics include observations for multiple assets. In contrast, the rest of the data consists of single time series. Here RI stands for total returns in percentages. The returns of individual stocks vary a lot more compared to the Fama–French portfolios. NYSE data includes many volatile stocks, and the 25 Fama–French portfolios are diversified. This high variance is also apparent in the market value (MV), price-to-book value ratio (PB), the price-to-earnings ratio (PE), price-to-cash flow ratio (PC), and dividend yield (DY) characteristics. These characteristics, apart from dividend yield, are not percentages, which also explains the high values. The total return is an average return of the sample and not a value-weighted market return. Stock returns are positively skewed, and a small number of extreme winners drive average returns.

I form NYSE factors with equal weighting. Baker and Wurgler (2006) also use the same methodology. The factors rely on NYSE data, which I explain more thoroughly in Section 3.1. The MV and PB factors should be close to the Fama–French factors SMB and HML. Interestingly, the returns differ from one another by a large margin. Different weighting or the volatility of individual companies may be the reason. The monthly mean of risk factor CMA is negative, indicating a negative premium from companies investing conservatively. The risk-free rate is low compared to its historical average.

The standard deviation of the sentiment indices in Table 1 shows a positive correlation between the number of observations and the standard deviation. This phenomenon also shows in Figure 5. I normalize the Google Trends data with the logarithmic difference before performing the principal component analysis. This step removes the first row of a data set, and the sentiment indices only have 167 observations. Lagged sentiment indices have 166 observations because delaying a time series removes the last observation. Table 2 shows the correlation matrix of the NYSE factors, Fama–French factors, and sentiment indices.

Table 2: Correlation matrices

This table presents the correlation matrices of the NYSE factors, Fama–French five factors, and sentiment indices. The factors are defined as described in Table 1.

NYSE factors

	MV SMB	PB Low−High	PE Low−High	PC Low−High	DY $>0-=0$
MV SMB	1.000	0.728	0.394	0.460	−0.697
PB Low−High	0.728	1.000	0.748	0.815	−0.484
PE Low−High	0.394	0.748	1.000	0.837	−0.324
PC Low−High	0.460	0.815	0.837	1.000	−0.433
DY $>0-=0$	−0.697	−0.484	−0.324	−0.433	1.000

Fama–French five factors

	$R_m - R_f$	SMB	HML	RMW	CMA
$R_m - R_f$	1.000	0.419	0.274	−0.463	−0.009
SMB	0.419	1.000	0.300	−0.402	0.149
HML	0.274	0.300	1.000	−0.185	0.474
RMW	−0.463	−0.402	−0.185	1.000	−0.065
CMA	−0.009	0.149	0.474	−0.065	1.000

Sentiment indices

	Sentiment 1	Sentiment 2	Sentiment 3	Sentiment 4
Sentiment 1	1.000	0.766	0.860	0.503
Sentiment 2	0.766	1.000	0.574	0.473
Sentiment 3	0.860	0.574	1.000	0.595
Sentiment 4	0.503	0.473	0.595	1.000

NYSE factors have a positive correlation, except for the dividend yield factor. I compute the dividend yield factor differently from the rest of the self-made factors, possibly explaining the result. Also, dividends can be more stable when compared to the earnings or cash flow, for example. Factors based on *PE* and *PC*, on the other hand, have a relatively high positive correlation. The Fama and French (1993) three-factor model factors are all positively correlated. In contrast, the new factors *RMW* and *CMA* of Fama and French (2015) negatively correlate with the excess market return. Factor *RMW* is also negatively correlated with the rest of the Fama–French factors.

Ideally, there would be little correlation among the factors to capture unique aspects of the data. In the Fama–French five-factor model, the absolute value of correlation does not exceed 0.5. High correlation among some of the NYSE factors could dilute their effectiveness when used together in a regression. The dividend yield factors should be more effective because it is uncorrelated with the rest of the NYSE factors.

The four sentiment indices have a positive correlation. I construct them from

the same data set, which can explain the correlation. The positive correlation can also indicate a common trend which the principal component analysis can extract. The first sentiment index includes all the search terms, whereas the second and third indices contain partially the same keywords. The fourth sentiment index only consists of one search query *debt*. It also has the least correlation with the other sentiment indices. The first sentient index has the most correlation with the other three. I do not present all the possible correlation combinations. However, the correlations between the sentiment indices and Fama–French factors could be appealing, for example.

The following Figure 6 and Table 3 demonstrate Fama–French portfolios' performance as explained in Section 3.2. The data consists of monthly percentage returns of 25 value-weighted portfolios formed on size and book-to-market.

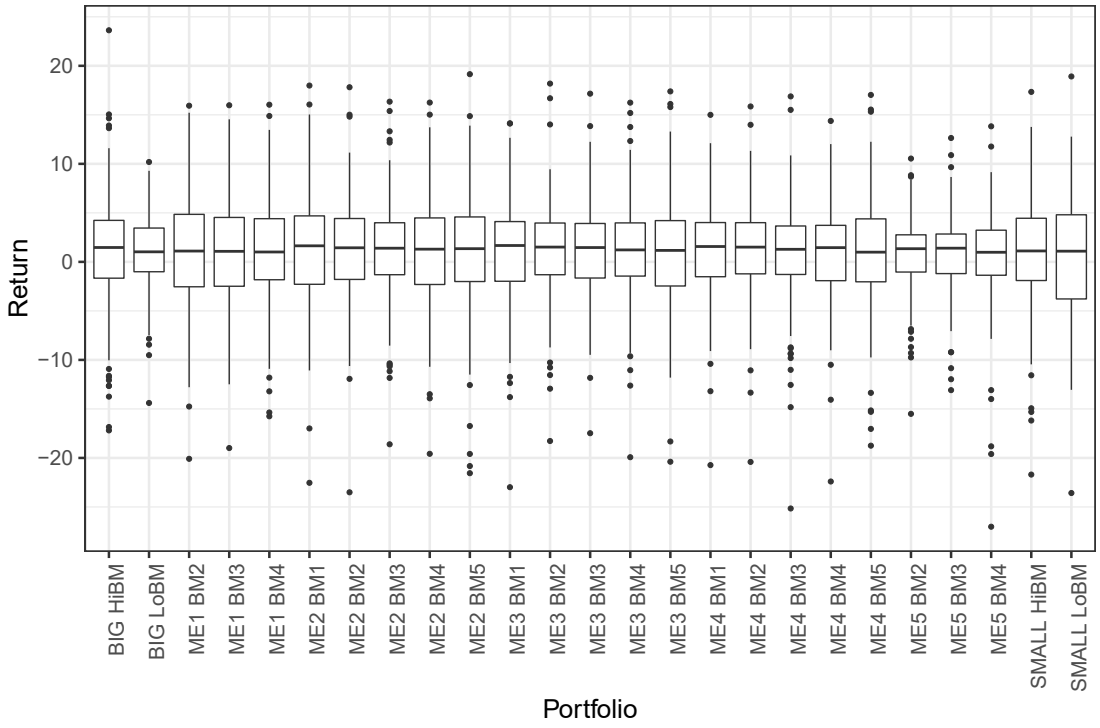

Figure 6: Fama–French 25 portfolio returns

All portfolios have a positive median return, which I expect given the positive market risk premium within the period. Portfolios with small market value and low book-to-market ratio appear to have higher average variance. However, Figure 6 does not indicate any clear pattern in the portfolio returns. The differences are apparent in Table 3, which presents the monthly mean percentage returns and standard deviations in a matrix form. Rows rank the portfolios based on book-

to-market characteristics from low to high on a scale of 1–5. Columns list the portfolios based on size from small to large on a scale of 1–5.

Table 3: Fama–French 25 portfolios

This table presents the returns of the Fama–French 25 portfolios formed on size and book-to-market. The portfolios are defined as described in Section 3.2. The number one denotes the portfolios with the smallest size or smallest book-to-market ratio. Consequently, number five denotes the portfolios with the largest size or highest book-to-market ratio.

Fama–French 25 portfolios: Mean

		Size				
		1	2	3	4	5
Book-to-market	1	0.391	0.946	0.895	1.047	0.818
	2	0.792	1.111	1.104	0.982	0.829
	3	0.758	1.087	1.067	0.811	0.832
	4	0.864	0.911	0.995	0.996	0.509
	5	0.919	0.875	0.994	0.891	0.954

Fama–French 25 portfolios: Standard deviation

		Size				
		1	2	3	4	5
Book-to-market	1	6.281	5.821	5.406	5.060	4.212
	2	5.885	5.560	5.284	5.035	4.166
	3	5.660	5.498	5.253	5.211	4.317
	4	5.482	5.343	5.208	5.035	4.754
	5	5.654	5.834	5.458	5.284	5.369

The small-cap growth stock portfolio has the highest volatility and the lowest average returns. This finding is in line with the literature where growth stocks with low market capitalization fail to deliver the size effect, diminishing the size factor (Fama and French 1993; 2015). Small size portfolios also appear to be a little more volatile on average, which I expect. Then again, a portfolio with the second-lowest size and the book-to-market ratio has the highest mean return with no apparent reason. The small-cap value stocks do not seem to outperform the rest of the portfolios, as Fama and French (1993) would suggest. There does not appear to be any clear pattern in the returns based on those two characteristics. However, regression models can identify possible size or value premiums more convincingly than these characteristics alone.

4.2 Fama–MacBeth regression results

Table 4 presents two Fama–MacBeth regressions with Fama–French five factors as right-hand variables. I explain the methodology in Section 3.5. I do not include sentiment factors in the regressions presented in this table. This way, I can see how the FF5 factors explain the cross-section of stock returns. Regression I uses NYSE stocks, and regression II uses Fama–French 25 portfolios as left-hand variables. The factor coefficients of a Fama–MacBeth regression represent an average risk premium the investor receives for exposure to the factor.

Table 4: Fama–MacBeth regressions without sentiment

This table presents Fama–MacBeth regressions on NYSE excess stock returns I, and Fama–MacBeth regressions on Fama–French excess portfolio returns II. I define the variables as described in Table 3 and Table 1. I present the t-statistics in parentheses under the coefficients.

Panel: Regressions I–II

	Excess return	
	(I)	(II)
Constant	1.018***	1.162**
	(4.388)	(2.507)
$\lambda_{R_m-R_f}$	0.349	0.153
	(1.017)	(0.331)
λ_{SMB}	0.374**	0.103
	(2.119)	(0.438)
λ_{HML}	0.226	−0.108
	(1.342)	(−0.417)
λ_{RMW}	−0.146	0.304
	(−1.494)	(1.355)
λ_{CMA}	0.050	−0.343
	(0.665)	(−1.604)
N	73,969	2,700

[1] ***p < 0.01, **p < 0.05, *p < 0.1

The intercepts in Table 4 are both significant at 0.01 level. Therefore, an investment generates a positive return on average when all factor returns are zero. According to the significance of the intercept, the right-hand variables do not explain the dependent factor well. In regression I, the lambda coefficient of the *SMB* factor is positive and significant at 0.05 level. The exposure to the size premium benefits the investor on average. However, this finding does not apply to regression II. According to the results, the size premium applies to average

returns of individual stocks and not widely diversified portfolios. The rest of the lambda coefficients are not significantly different from zero.

I find right-hand variables to be surprisingly insignificant. As explained in Section 3.5, I use a time window of 60 months for rolling regressions. Due to NYSE data limitations, an asset must have at least 45 out of 60 valid observations for regressions. These choices might impact the results of regression I but not regression II. Therefore, the factors might be less significant during the sample period than the literature suggests.

The following Table 5 presents the results of Fama–MacBeth regressions on NYSE stocks with sentiment indices. Each of the four regressions includes a unique sentiment factor.

Table 5: Fama–MacBeth regressions on NYSE stocks

This table presents Fama–MacBeth regressions on NYSE stocks I–IV. I define the variables as described in Table 1. I present the t-statistics in parentheses under the coefficients.

Panel: Regressions I–IV

	Excess return			
	(I)	(II)	(III)	(IV)
Constant	1.014***	1.034***	1.007***	1.006***
	(4.342)	(4.425)	(4.326)	(4.324)
$\lambda_{\text{Sentiment 1}}$	0.043			
	(0.572)			
$\lambda_{\text{Sentiment 2}}$		0.047		
		(0.629)		
$\lambda_{\text{Sentiment 3}}$			0.045	
			(0.595)	
$\lambda_{\text{Sentiment 4}}$				0.041
				(0.560)
$\lambda_{\text{R}_\text{m}-\text{R}_\text{f}}$	0.014	−0.023	−0.007	−0.016*
	(0.374)	(−1.101)	(−0.302)	(−1.675)
λ_{SMB}	0.368	0.356	0.374	0.367
	(1.068)	(1.033)	(1.087)	(1.069)
λ_{HML}	0.362**	0.365**	0.361**	0.369**
	(2.059)	(2.075)	(2.050)	(2.102)
λ_{RMW}	0.209	0.217	0.215	0.228
	(1.244)	(1.291)	(1.273)	(1.360)
λ_{CMA}	−0.136	−0.142	−0.135	−0.141
	(−1.383)	(−1.439)	(−1.384)	(−1.448)
N	73,969	73,969	73,969	73,969

[1] ***p < 0.01, **p < 0.05, *p < 0.1

The intercepts' coefficients in each of the four Fama–MacBeth regressions are positive and significantly different from zero at 0.01 significance level. However, I am more interested in the lambda coefficients' significance, and especially in the sentiment indices' lambdas. The lambdas of the four sentiment indices are also positive, but none of them are significant. According to the t-statistics of the Fama–MacBeth regressions, there is no compensation from the exposure to the Google Trends-based sentiment factors. The risk premiums are not significant even at a 90% confidence level. I accept the null hypothesis, which I present in Equation (23), of lambda coefficients being equal to zero. Accordingly, the Google Trends-based sentiment index does not explain the cross-section of asset returns, which answers the second research question.

I find the Fama–French five factors to be oddly insignificant. As opposed to the results presented in Table 4, now the *HML* factor has become significant. Also, the *SMB* factor has lost its significance. The sentiment indices may capture some information impacting the Fama–French five factors. The *HML* factor is positive and significant at a 0.05 level across all four regressions. Thus, exposure to the *HML* generates a risk premium at a 95% confidence level.

The rest of the lambda coefficients are not statistically significant, apart from the excess market return factor, which is significant in regression IV at the 0.1 significance level. Interestingly, this excess market return factor becomes negative when including the sentiment variable. However, this sign flipping of the coefficient is due to the correlation of left-hand and right-hand variables. Intuitively, both excess returns of individual stocks and the market return factor reflect the same information, which causes bias in the estimator.

Table 6 presents the results of Fama–MacBeth regressions on Fama–French 25 portfolios. Regressions I–IV include the four sentiment indices, whereas regressions V–VII contain lagged regressions. Each of the eight regressions consists of a unique sentiment factor.

Table 6: Fama–MacBeth regressions on Fama–French 25 portfolios

This table presents the results of Fama–MacBeth regressions on Fama–French 25 portfolios *I–VIII*. I define the variables as described in Table 3 and Table 1. I present the *t*-statistics in parentheses under the coefficients.

Panel: Regressions I–IV

	(I)	(II)	(III)	(IV)
Constant	1.263***	1.226**	1.214**	1.183**
	(2.652)	(2.472)	(2.553)	(2.463)
$\lambda_{\text{Sentiment 1}}$	0.035			
	(0.267)			
$\lambda_{\text{Sentiment 2}}$		0.014		
		(0.224)		
$\lambda_{\text{Sentiment 3}}$			0.030	
			(0.357)	
$\lambda_{\text{Sentiment 4}}$				0.014
				(0.432)
$\lambda_{\text{R}_\text{m}-\text{R}_\text{f}}$	0.143	0.181	0.196	0.225
	(0.290)	(0.355)	(0.384)	(0.450)
λ_{SMB}	0.118	0.123	0.114	0.112
	(0.501)	(0.521)	(0.481)	(0.474)
λ_{HML}	−0.009	−0.006	−0.020	−0.005
	(−0.035)	(−0.026)	(−0.081)	(−0.021)
λ_{RMW}	0.303	0.303	0.283	0.311
	(1.361)	(1.384)	(1.279)	(1.421)
λ_{CMA}	−0.315	−0.287	−0.212	−0.351*
	(−1.488)	(−1.334)	(−1.021)	(−1.687)
N	2,675	2,675	2,675	2,675

Panel: Regressions V–VIII

	(V)	(VI)	(VII)	(VIII)
Constant	1.203**	1.401***	1.233***	1.193**
	(2.459)	(2.948)	(2.606)	(2.554)
$\lambda_{\text{Sentiment 1 (−1)}}$	−0.012			
	(−0.078)			
$\lambda_{\text{Sentiment 2 (−1)}}$		−0.063		
		(−0.928)		
$\lambda_{\text{Sentiment 3 (−1)}}$			0.005	
			(0.048)	
$\lambda_{\text{Sentiment 4 (−1)}}$				0.004
				(0.119)
$\lambda_{\text{R}_\text{m}-\text{R}_\text{f}}$	0.307	0.102	0.283	0.318
	(0.622)	(0.214)	(0.612)	(0.680)
λ_{SMB}	0.138	0.136	0.129	0.124
	(0.579)	(0.571)	(0.541)	(0.521)
λ_{HML}	0.054	0.066	0.055	0.067
	(0.230)	(0.283)	(0.235)	(0.286)
λ_{RMW}	0.287	0.320	0.279	0.223
	(1.266)	(1.409)	(1.239)	(0.967)
λ_{CMA}	−0.299	−0.322	−0.254	−0.268
	(−1.388)	(−1.497)	(−1.193)	(−1.306)
N	2,650	2,650	2,650	2,650

[1] ***p < 0.01, **p < 0.05, *p < 0.1

The Fama–French 25 portfolios as left-hand variables instead of NYSE stocks results in similar, but even less statistically significant, results. The intercepts in all the regressions are significant, at least at a 0.05 significance level. The constants in regressions I, VI, VII, $VIII$ are significant at 0.01 level. Sentiment indices and FF5 factors overall have little statistical significance. The only significant factor is *CMA* in regression IV at a 90% confidence level. The sentiment indices do not significantly improve the Fama–French five-factor model.

Lagged sentiments do not explain the cross-section of asset return any better. With lagged sentiments, the sentiment coefficients become negative in regressions V and VI. Also, *HML* coefficients become positive when using lagged sentiments. However, the findings become meaningless due to low t-statistics and coefficients being indifferent from zero. I use a lag of one as explained in Section 3.4. Intuitively, it does not make sense to use a more significant lag since I use monthly data. Google Trends information unlikely explains the asset returns over a month ahead.

According to the results, the Google Trends-based sentiment does not explain the cross-section of asset returns. The results in Table 5 support the findings in Table 6. These results confirm the null hypothesis of lambda coefficients being equal to zero.

4.3 Results from other regression-based tests

I test various clustering methods using excess NYSE stock returns as left-hand variables and present the results in Table 7. The right-hand variables are the sentiment indices and NYSE factors. I use clustering by a firm in regressions I–IV, clustering by month in regressions V–$VIII$, and double clustering by firm and month in regressions IX–$XIII$.

Table 7: Regressions on NYSE stocks with different clustering

This table presents the results of regressions on NYSE stocks with different clustering I–XII. I define the variables as described in Table 1. I present the t-statistics in parentheses under the coefficients. I use White standard errors are, N is 143,609 in all regressions. Also, Firm (F) and Month (M) on the clustering row denote firm and month clustering, respectively.

Panel A: Regressions I–VI

	(I)	(II)	(III)	(IV)	(V)	(VI)
Constant	2.353***	2.320***	2.353***	2.348***	2.353***	2.320***
	(50.001)	(49.383)	(49.989)	(49.808)	(6.453)	(6.580)
Sentiment 1	−0.137***				−0.137	
	(−4.167)				(−0.409)	
Sentiment 2		−1.453***				−1.453*
		(−18.605)				(−1.754)
Sentiment 3			0.280***			
			(4.897)			
Sentiment 4				−0.299*		
				(−1.905)		
MV SMB	0.265***	0.282***	0.271***	0.266***	0.265	0.282*
	(10.859)	(11.611)	(11.104)	(10.930)	(1.558)	(1.682)
PB Low−High	−0.032	0.007	−0.043	−0.040	−0.032	0.007
	(−0.971)	(0.218)	(−1.311)	(−1.213)	(−0.115)	(0.027)
PE Low−High	0.369***	0.316***	0.383***	0.368***	0.369*	0.316*
	(15.345)	(13.229)	(15.793)	(15.121)	(1.887)	(1.653)
PC Low−High	0.436***	0.428***	0.427***	0.444***	0.436**	0.428**
	(11.949)	(11.710)	(11.651)	(12.116)	(2.458)	(2.470)
DY $>0-=0$	−1.090***	−1.026***	−1.103***	−1.094***	−1.090***	−1.026***
	(−32.234)	(−30.386)	(−32.658)	(−32.432)	(−4.630)	(−4.323)
Clustering	Firm	Firm	Firm	Firm	Month	Month
Adjusted R^2	0.141	0.143	0.141	0.141	0.141	0.143

Panel B: Regressions VII–XII

	(VII)	(VIII)	(IX)	(X)	(XI)	(XII)
Constant	2.353***	2.348***	2.353***	2.320***	2.353***	2.348***
	(6.379)	(6.369)	(6.445)	(6.572)	(6.372)	(6.362)
Sentiment 1			−0.137			
			(−0.409)			
Sentiment 2				−1.453*		
				(−1.755)		
Sentiment 3	0.280				0.280	
	(0.494)				(0.494)	
Sentiment 4		−0.299				−0.299
		(−0.177)				(−0.177)
MV SMB	0.271	0.266	0.265	0.282*	0.271	0.266
	(1.597)	(1.564)	(1.550)	(1.673)	(1.589)	(1.556)
PB Low−High	−0.043	−0.040	−0.032	0.007	−0.043	−0.040
	(−0.151)	(−0.141)	(−0.114)	(0.027)	(−0.151)	(−0.140)
PE Low−High	0.383*	0.368*	0.369*	0.316*	0.383*	0.368*
	(1.957)	(1.840)	(1.885)	(1.651)	(1.955)	(1.839)
PC Low−High	0.427**	0.444**	0.436**	0.428**	0.427**	0.444**
	(2.394)	(2.462)	(2.427)	(2.436)	(2.363)	(2.431)
DY $>0-=0$	−1.103***	−1.094***	−1.090***	−1.026***	−1.103***	−1.094***
	(−4.708)	(−4.696)	(−4.604)	(−4.300)	(−4.681)	(−4.669)
Clustering	Month	Month	F/M	F/M	F/M	F/M
Adjusted R^2	0.141	0.141	0.141	0.143	0.141	0.141

[1] $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Clustering by firm appears to generate significant results across the factors. However, this is an incorrect clustering method for my panel data. This method generates, on average, over ten times greater t-statistics than the other clustering methods. Clustering by month and double clustering generate practically identical results. I do not include results without clustering, but the t-statistics of those regressions would be greater than or equal to those with firm clustering. This finding indicates clustering to be necessary. I focus on the double clustered regression results from now on.

Double clustered regressions X–XII have significant intercepts at a 0.01 significance level. The second sentiment index is also significant in regression X but only at the 0.1 level. The coefficient is negative at -1.453, which implies sentiment risk to be unattractive to investors on average. In practice, the high level of sentiment would indicate lower subsequent returns, which is in line with the prevailing literature. However, the factor is significant only at 0.1 level, and the other sentiment factors in double clustered regressions are not significant even at a 90% confidence level.

The factor based on DY is significant across the regressions. The coefficients are consistently negative, implying a negative equity risk premium among dividend-paying stocks. I expect this factor to be significant given its small correlation with other factors. However, I am somewhat surprised by the negative coefficient. Factors based on PC and PE imply positive risk equity risk premiums in all double clustered regressions. The factors based on the price-to-cash flow ratio are significant at 0.05, and factors based on the price-to-earnings rate are significant at 0.1. The factors PC and PE use common value stock characteristics. The results imply positive equity risk premiums for value stocks. However, factors based on other typical value stock ratios, such as market value, and price-to-book value ratio, are not significant. The exception is the market value-based ratio MV in regression X with a 0.1 significance level.

In Table 8, I use the self-constructed cross-sectional NYSE factors as left-hand variables in regressions. These are essentially long-short portfolios. I also include excess stock return as a left-hand variable for reference. I use sentiment indices individually as right-hand variables. This regression method attempts to explain the returns of cross-sectional portfolios based on different characteristics with sentiment factors. The regressions test if the synthetic sentiment indices constructed in this thesis are related to aggregate stock market movements. In Table 8, I use sentiment as the only factor in a model.

Table 8: Long-short regressions

This table presents the results of long-short regressions I–$XXIV$. I define the variables as described in Table 1. I present the t-statistics in parentheses under the coefficients and use White standard errors and firm and month clustering in all regressions. Also, N is 143,609 in all regressions, respectively.

Panel A: Regressions I–VI

	Excess return	MV SMB	PB Low−High	PE Low−High	PC Low−High	DY $>0-=0$
	(I)	(II)	(II)	(IV)	(V)	(VI)
Constant	1.005***	−0.334	−1.910***	−2.282***	−1.585***	−0.195
	(2.762)	(−1.437)	(−10.852)	(−12.717)	(−7.769)	(−1.430)
Sentiment 1	−0.780	−0.348	−0.139	−0.241	−0.260	0.324
	(−1.283)	(−1.054)	(−0.529)	(−0.724)	(−0.724)	(1.396)
Adjusted R^2	0.003	0.007	0.002	0.006	0.005	0.018

Panel B: Regressions VII–XII

	Excess return	MV SMB	PB Low−High	PE Low−High	PC Low−High	DY $>0-=0$
	(VII)	(VIII)	(IX)	(X)	(XI)	(XII)
Constant	1.026***	−0.329	−1.906***	−2.275***	−1.578***	−0.202
	(2.913)	(−1.415)	(−10.845)	(−12.907)	(−7.801)	(−1.496)
Sentiment 2	−3.586**	−0.533	−0.579	−1.325	−1.282	0.986
	(−2.304)	(−0.579)	(−0.754)	(−1.521)	(−1.346)	(1.615)
Adjusted R^2	0.015	0.004	0.008	0.038	0.028	0.037

Panel C: Regressions XIII–XVIII

	Excess return	MV SMB	PB Low−High	PE Low−High	PC Low−High	DY $>0-=0$
	(XII)	(XIV)	(XV)	(XVI)	(XVII)	(XVIII)
Constant	1.011***	−0.334	−1.909***	−2.281***	−1.582***	−0.196
	(2.761)	(−1.440)	(−10.819)	(−12.677)	(−7.731)	(−1.436)
Sentiment 3	−0.363	−0.723	−0.101	−0.139	0.204	0.440
	(−0.377)	(−1.351)	(−0.227)	(−0.270)	(0.354)	(1.230)
Adjusted R^2	0.001	0.007	0.001	0.004	0.000	0.008

Panel D: Regressions XIX–XXIV

	Excess return	MV SMB	PB Low−High	PE Low−High	PC Low−High	DY $>0-=0$
	(XIX)	(XX)	(XXI)	(XXII)	(XIII)	(XXIV)
Constant	1.011***	−0.333	−1.909***	−2.281***	−1.583***	−0.197
	(2.775)	(−1.431)	(−10.838)	(−12.722)	(−7.743)	(−1.439)
Sentiment 4	−2.126	−1.658	−0.488	−1.035	0.258	1.041
	(−0.855)	(−1.461)	(−0.523)	(−1.001)	(0.208)	(1.237)
Adjusted R^2	0.001	0.007	0.001	0.004	0.000	0.008

[1] ***p < 0.01, **p < 0.05, *p < 0.1

Based on the overall regression results, the synthetic sentiment indices constructed in the thesis are not related to aggregate stock market movements. These findings provide evidence for the first research question. In regressions where I use long-short portfolios as dependent variables, the sentiment factors are not significant. Based on this result, portfolios formed on stock-specific characteristics are not sensitive to the Google Trends-based sentiment changes. In regression VII, the coefficient of the second sentiment index is significant at 0.05 level. This sentiment is the only significant factor in Table 8. The negative coefficient implies exposure to this sentiment factor to be unfavorable.

The results align with the findings of Preis et al. (2013). They take a short position when the search volume increases and vice versa. The results of regression VII indicate a negative relationship between the sentiment and the returns. Being short when the sentiment is high and being long when it is low would be beneficial, as Preis et al. (2013) suggest. The second sentiment index, which consists of the 15 best-performing search queries of Preis et al. (2013), is the only significant factor. This finding indicates these search terms to be somewhat related to the aggregate stock market movements. However, one significant sentiment coefficient in 24 different regressions does not support this finding. I find the evidence insufficient to reject the null hypothesis.

The constants are all significant at 0.01 level when the left-hand variable is based on excess return, *PB*, *PE*, or *PC*. However, the constants are not significant when the left-hand variable is based on *MV* or *DY*. Sentiment indices alone do not explain the dependent variable as close-to-zero adjusted R^2 values indicate. The constants of the factors based on *PB*, *PE*, and *DY* are likely negative because, in Table 1, the means of those factors are highly negative. The means of factors constructed on *MV* and *PC* are close to zero. When I use those factors as left-hand variables, the constants are insignificant as well. On the contrary, the excess return has a highly positive mean due to positive aggregate stock returns and the low risk-free rates. The intercepts in those regressions are statistically significant.

In Table 9 and Table 10, I use the same methodology as in Table 8. However, I add Fama–French five factors as right-hand variables. I split the results into two tables to keep the results legible. I exclude the *SMB* factor as a control variable when the long-short portfolio based on *MV* is the dependent variable. I also exclude the *HML* factor as a control variable when the long-short portfolio based on *PB* is the dependent variable. These factors or portfolios are based on the same criteria so that the fit may be too perfect. This similarity can then disrupt the other factors.

Table 9: Long-short regressions with Fama–French five factors 1/2

This table presents the results of long-short regressions with Fama–French five factors *I–XII*. I define the variables as described in Table 1 and present the *t*-statistics in parentheses under the coefficients. I use White standard errors and double clustering by firm and month in all regressions. Also, N is 143,609 in all regressions, respectively.

Panel A: Regressions I–VI

	Excess return	MV SMB	PB Low−High	PE Low−High	PC Low−High	DY $>0-=0$
	(I)	(II)	(III)	(IV)	(V)	(VI)
Constant	0.134	−0.384**	−2.088***	−2.495***	−1.830***	−0.112
	(1.371)	(−2.248)	(−13.898)	(−15.106)	(−9.901)	(−1.339)
Sentiment 1	0.044	−0.212	0.107	−0.155	−0.134	0.082
	(0.307)	(−0.967)	(0.517)	(−0.594)	(−0.490)	(0.751)
$R_m - R_f$	1.084***	0.238***	0.261***	0.269***	0.279***	−0.134***
	(32.068)	(4.450)	(4.631)	(4.162)	(4.342)	(−5.478)
SMB	0.445***		0.223***	−0.130*	−0.003	−0.395***
	(8.611)		(2.848)	(−1.664)	(−0.035)	(−8.970)
HML	0.031	0.504***		0.392***	0.480***	−0.017
	(0.506)	(4.806)		(4.578)	(3.863)	(−0.359)
RMW	0.206***	−0.364***	−0.037	0.049	0.105	0.175***
	(3.039)	(−3.284)	(−0.322)	(0.441)	(0.808)	(2.848)
CMA	−0.090	0.039	0.274**	−0.414***	−0.380**	0.262***
	(−1.215)	(0.247)	(2.572)	(−2.970)	(−2.261)	(3.337)
Adjusted R^2	0.215	0.470	0.384	0.375	0.400	0.660

Panel B: Regressions VII–XII

	Excess return	MV SMB	PB Low−High	PE Low−High	PC Low−High	DY $>0-=0$
	(VII)	(VIII)	(IX)	(X)	(XI)	(XII)
Constant	0.139	−0.389**	−2.091***	−2.481***	−1.816***	−0.118
	(1.413)	(−2.276)	(−13.915)	(−15.321)	(−9.876)	(−1.432)
Sentiment 2	−0.176	0.070	0.190	−0.643	−0.630	0.283
	(−0.533)	(0.141)	(0.357)	(−1.057)	(−0.965)	(0.975)
$R_m - R_f$	1.078***	0.246***	0.264***	0.253***	0.263***	−0.127***
	(32.131)	(4.609)	(4.815)	(4.153)	(4.286)	(−5.005)
SMB	0.444***		0.218***	−0.121	0.005	−0.399***
	(8.802)		(2.842)	(−1.620)	(0.068)	(−9.007)
HML	0.036	0.498***		0.400***	0.488***	−0.020
	(0.592)	(4.843)		(4.810)	(4.024)	(−0.426)
RMW	0.204***	−0.362***	−0.038	0.051	0.106	0.174***
	(3.041)	(−3.295)	(−0.334)	(0.460)	(0.816)	(2.817)
CMA	−0.089	0.043	0.270**	−0.404***	−0.371**	0.257***
	(−1.217)	(0.270)	(2.458)	(−2.863)	(−2.145)	(3.245)
Adjusted R^2	0.215	0.467	0.384	0.381	0.405	0.662

[1] ***p < 0.01, **p < 0.05, *p < 0.1

[2] Control variable *SMB* (*HML*) is not included when *MV* (*PB*) is the dependent variable.

Table 10: Long-short regressions with Fama–French five factors 2/2

This table presents the results of long-short regressions with Fama–French five factors *XIII*–*XXIV*. I define the variables as described in Table 1 and present the *t*-statistics in parentheses under the coefficients. I use White standard errors and double clustering by firm and month in all regressions. Also, N is 143,609 in all regressions, respectively.

Panel C: Regressions XIII–XVIII

	Excess return	MV SMB	PB Low−High	PE Low−High	PC Low−High	DY $>0-=0$
	(XIII)	(XIV)	(XV)	(XVI)	(XVII)	(XVIII)
Constant	0.136	−0.393**	−2.085***	−2.499***	−1.830***	−0.110
	(1.390)	(−2.298)	(−13.815)	(−15.033)	(−9.846)	(−1.320)
Sentiment 3	0.248	−0.573	0.207	−0.259	0.204	0.029
	(1.087)	(−1.549)	(0.567)	(−0.628)	(0.457)	(0.162)
$R_m - R_f$	1.082***	0.243***	0.258***	0.273***	0.281***	−0.135***
	(30.459)	(4.519)	(4.465)	(4.104)	(4.045)	(−5.535)
SMB	0.452***		0.227***	−0.134*	0.008	−0.396***
	(8.670)		(2.863)	(−1.670)	(0.101)	(−8.859)
HML	0.029	0.504***		0.391***	0.474***	−0.015
	(0.474)	(4.749)		(4.435)	(3.780)	(−0.319)
RMW	0.203***	−0.350***	−0.040	0.054	0.106	0.173***
	(3.077)	(−3.144)	(−0.352)	(0.488)	(0.831)	(2.806)
CMA	−0.088	0.034	0.275***	−0.414***	−0.375**	0.261***
	(−1.186)	(0.214)	(2.593)	(−2.971)	(−2.267)	(3.340)
Adjusted R^2	0.215	0.474	0.384	0.375	0.400	0.659

Panel D: Regressions XIX–XXIV

	Excess return	MV SMB	PB Low−High	PE Low−High	PC Low−High	DY $>0-=0$
	(XIX)	(XX)	(XXI)	(XXII)	(XIII)	(XXIV)
Constant	0.136	−0.388**	−2.085***	−2.499***	−1.829***	−0.110
	(1.386)	(−2.256)	(−13.849)	(−15.002)	(−9.889)	(−1.317)
Sentiment 4	0.398	−0.386	0.465	−0.608	0.887	0.095
	(0.742)	(−0.397)	(0.605)	(−0.728)	(0.880)	(0.201)
$R_m - R_f$	1.084***	0.243***	0.259***	0.272***	0.283***	−0.135***
	(31.294)	(4.604)	(4.564)	(4.077)	(4.157)	(−5.530)
SMB	0.444***		0.221***	−0.127*	0.003	−0.397***
	(8.687)		(2.857)	(−1.652)	(0.039)	(−8.921)
HML	0.032	0.500***		0.389***	0.476***	−0.015
	(0.517)	(4.858)		(4.329)	(3.742)	(−0.312)
RMW	0.201***	−0.359***	−0.043	0.058	0.099	0.172***
	(3.011)	(−3.240)	(−0.368)	(0.523)	(0.769)	(2.771)
CMA	−0.090	0.042	0.276***	−0.413***	−0.374**	0.261***
	(−1.209)	(0.264)	(2.601)	(−2.957)	(−2.258)	(3.332)
Adjusted R^2	0.215	0.468	0.384	0.374	0.401	0.659

[1] ***p < 0.01, **p < 0.05, *p < 0.1

[2] Control variable *SMB* (*HML*) is not included when *MV* (*PB*) is the dependent variable.

The sentiment indices cannot explain the aggregate stock returns and do not complement the Fama–French five-factor model. In Table 9 and Table 10, none of the four sentiment indices are significant. Even the second sentiment index, which is significant in Table 8 when the excess return is the dependent variable, is not significant. When I include the Fama–French factors as control variables, the model explains the left-hand variable better, as the higher adjusted R^2 values indicate. The intercept in regressions in which the dependent variable is excess return is close to zero. This result shows the five factors to explain most of the expected stock returns. However, I am not particularly interested in the constants since they are not the focus of this study.

Factor $R_m - R_f$ has t-statistics above 30 when the left-hand variable is the excess stock return. This factor is significant at 0.01 level across all regressions, which implies the existence of equity risk premium. This significance makes sense because the left-hand variable represents an excess return for individual stocks. The right-hand variable represents an excess return for portfolios. The other Fama–French five factors are significant with some dependent variables but insignificant with others. For example, when the dependent variable based on DY, the factor HML is not significant. In contrast, the other Fama–French factors are significant at 0.01 level. Therefore, the value premium factor does not explain the difference in returns between dividend-paying and non-dividend-paying stocks. However, the other Fama–French factors do. There appears to be a connection between size and value. For example, the value premium factor is significant at 0.01 level when the dependent variable bases on market value. The size premium factor is also significant at 0.01 level when the dependent factor relies on book value. These observations are interesting but not the focus of this thesis.

Table 11 shows results from risk-adjusted regressions on NYSE stocks. In this modified version of the Fama–MacBeth regression, I first calculate the beta estimators with Fama–French five-factor model. Then I multiply the factors with the corresponding beta estimators, as explained in Section 3.5. In the second step, I use the beta estimators of sentiment indices as independent variables. I must use the beta estimators because the sentiment indices do not have asset-specific values. Risk-adjusted regressions face the same issues as the Fama–MacBeth regressions do. The rolling time window of rolling regressions and the problems in NYSE data may impact the results. NYSE factors and Fama–French factors related to size and value may capture similar risk, impacting results.

Table 11: Risk-adjusted regressions on NYSE stocks

This table presents the results of risk-adjusted regressions on NYSE stocks *I–IV*. I define the variables as described in Table 1 and present the *t*-statistics in parentheses under the coefficients.

Panel: Regressions I–IV

	Risk-adjusted return			
	(I)	(II)	(III)	(IV)
Constant	1.300***	1.648***	1.212***	1.992***
	(3.344)	(3.096)	(3.096)	(2.843)
$\lambda_{\text{Sentiment 1}}$	0.073			
	(0.634)			
$\lambda_{\text{Sentiment 2}}$		0.174		
		(1.194)		
$\lambda_{\text{Sentiment 3}}$			0.209	
			(1.636)	
$\lambda_{\text{Sentiment 4}}$				−0.098
				(−0.387)
β_{MV}	0.000	0.000	0.000	0.000*
	(0.501)	(0.109)	(0.442)	(−1.665)
β_{PB}	0.005*	0.007*	0.004	0.009
	(1.943)	(1.873)	(0.729)	(0.758)
β_{PE}	0.009**	0.010**	0.009*	0.017
	(2.397)	(2.436)	(1.971)	(1.075)
β_{PC}	0.010**	0.003	0.009	−0.003
	(2.178)	(0.346)	(1.541)	(−0.160)
β_{DY}	−0.157***	−0.226***	−0.209***	−0.320***
	(−3.983)	(−5.036)	(−4.753)	(−2.769)
N	73,969	73,969	73,969	73,969
Adjusted R^2	0.003	0.001	0.002	0.000

[1] ***p < 0.01, **p < 0.05, *p < 0.1

The sentiment indices are not significantly different from zero. Therefore, the sentiments cannot explain the risk-adjusted returns. The third sentiment is the closest to being significant with *t*-statistics of 1.636. The intercepts are all significant and positive at a 0.01 significance level. The model using sentiment and self-constructed cross-sectional factors does not capture risk-adjusted returns well based on the intercepts. Adjusted R^2 values are also close to zero.

NYSE factor β_{DY} is significant in all four regression at a 0.01 significance level. Also, the negative coefficients indicate dividend-paying stocks having negative risk premiums. Factor β_{PE} is significant and positive in regressions *I* and *II* at 0.05 significance level and 0.1 in regression *III*. Other self-constructed factors are significant in some regressions but not significant in others.

5 SUMMARY AND CONCLUSION

5.1 Summary of findings

This thesis set off to explain the cross-section of stock asset returns with Google Trends using methods from multiple studies. Preis et al. (2013) report impressive results in their research, but their approach may not apply to asset pricing. The principal component analysis approach suggested by Baker and Wurgler (2006) enables to construct sentiment indices and test the keywords of Preis et al. (2013) in an asset pricing context. The five factors of Fama and French (2015) offer a leading-practice benchmark for testing the sentiment indices. Regression methods, such as the procedure of Fama and MacBeth (1973), enable empirical testing.

Based on the results, the synthetic sentiment indices constructed in this thesis do not relate to aggregate stock market movements. Based on the t-statistics of regressions, the sentiment factors are not significant when applying appropriate clustering and standard error corrections methods. Also, the Google Trends-based sentiment does not explain the cross-section of asset returns. The t-statistics of the Fama–MacBeth regressions do not indicate any compensation from the exposure to the Google Trends-based sentiment factors. The risk premiums are not significant even at a 90% confidence level. I accept the null hypothesis of risk premium estimators being indifferent from zero. The findings are not in line with the Preis et al. (2013), which is understandable due to the methodological differences between our studies.

The sentiment indices are not significant in any of the Fama–MacBeth regressions when I include Fama–French five factors in the model. When I use excess stock returns as the left-hand variables, the t-statistics vary between 0.560 and 0.629. When I use Fama–French 25 portfolios as left-hand variables instead, the t-statistics range from 0.224 to 0.432. Lagged sentiments generate even less significant results. In double clustered regressions, only one of the four sentiment indices is significant at a 90% confidence level. The coefficient is negative at -1.453, which implies sentiment risk to be unattractive to investors on average. Also, the second sentiment index is significant in only one of the long-short regressions. It is meaningful when I use excess stock return as the left-hand variable and the sentiment factor as the only right-hand variable. The coefficient value -3.586 is significant at the 0.05 level. However, the sentiment is insignificant in any other regressions. These findings are in line with Chordia et al. (2017) because they argue many

trading strategy studies test the hypothesis incorrectly and therefore report overly significant results.

The methodological ambiguity of the study of Preis et al. (2013) created a need to replicate the research using the same search terms but in a more appropriate empirical setting. Having done that, the explanatory power of the search terms in aggregate is nonexistent, unlike the initial study indicates. Intuitively, this result makes sense. Information extracted from a sample of various search terms is unlikely to explain the cross-section of stock returns. Also, any possible impact of the internet search activity is likely not visible due to the monthly frequency of the data. The prevailing literature suggests a negative relationship between the sentiment and the returns, where a high sentiment level would indicate lower subsequent returns. Based on the results, I can neither confirm nor deny this claim with any confidence.

5.2 Contribution to prior literature

The contribution of this thesis to the prior asset pricing literature is three-fold. I introduce a novel approach to construct a market sentiment index using principal component analysis (PCA) and data from Google Trends. Then, I show that a market sentiment index composed this way does not explain the cross-section of stock returns using rigorous asset pricing tests. Finally, I run multiple robustness checks using alternative methods to verify the central results and discuss the possible issues of the approach.

In this study, I combine the non-asset pricing Google Trends study of Preis et al. (2013) and a non-Google Tends-related asset pricing study of Baker and Wurgler (2006). This thesis contributes to Google Trends-related research, in general, as it evaluates the usefulness of Google Trends data in theoretical applications. The thesis also contributes to a particular segment of asset pricing, studying the explanatory capabilities of alternative factors in multi-factor models. The availability and the amount of Google Trends data make it an appealing source for sentiment proxies. However, it may not be as beneficial of a predictor as previous studies indicate. Overall, the empirical evidence does not support using Google Trends-based factors in asset pricing models. This finding is in line with the Fama–Frech framework (see, e.g., Fama and French 1993; 2015).

I demonstrate PCA to be a practical tool in creating new asset pricing factors from practically any data set. The steps described in Section 3.4 and the R code in

Appendix 1 offer a structure for the upcoming empirical studies. Prior literature suggests Google Trends data having many practical applications. For example, Ginsberg et al. (2009) demonstrate how the data may detect influenza epidemics. Choi and Varian (2012) forecast different near-term economic indicators before the release of the official figures. Using the demonstrated novel approach to construct a market sentiment index allows testing all sorts of data in different multi-factor regression models.

Based on the empirical results of this thesis, the Google Trends-based sentiment does not explain the cross-section of asset returns. I use the procedure of Fama and MacBeth (1973) to test the cross-sectional significance. This thesis contributes to the respective literature and demonstrates the method being useful also in Google Trends-based applications. In Fama–MacBeth regressions, I use the factors of Fama and French (2015) to test the sentiment indices in a multi-factor setting. Therefore, the thesis also contributes to the asset pricing literature in general. New factors and approaches may strengthen the understanding of the expected asset returns. Regardless, the literature and the empirical evidence of this thesis support the Fama–French five-factor model as a leading practice in asset pricing. However, research on alternative factors may identify new risk premiums and complement the existing models.

The alternative regression methods verify that the synthetic sentiment indices constructed in this thesis do not relate to aggregate stock market movements. The results are in line with the ones from the Fama and MacBeth (1973) procedure. The alternative methods in question include clustering suggested by Cochrane (2005, 245–252) and Petersen (2009), so-called long-short regressions from Baker and Wurgler (2006), and risk-adjusted regressions inspired by Brennan et al. (1998). These robustness checks, commonly used in prior literature, enable the verification of the central findings. Therefore, this thesis also contributes to the respective literature and practical implications of these methods. According to my understanding, this thesis is the first study applying the approaches to my research questions, and the robustness checks appear to fit the purpose. Additional methods are unlikely required because the risk premiums are not significant after the above-mentioned robust checks. However, the issues and limitations of the data may disrupt the results and make it difficult to evaluate the effectiveness of the approach.

This thesis evaluates the usefulness of Google Trends data in theoretical applications. Even if significant, Google Trends-based sentiment is not an investable factor. Therefore, it would have few real-world implications in the construction of investment portfolios, for example. From a practical standpoint, the literature

implies investing in a portfolio with high FF5 factor loadings. In reality, such an investment vehicle may not be available. Also, constructing a comparable portfolio from individual securities would be both difficult and expensive. Investable factor funds may differ from the portfolios used in academic studies (see, e.g., Fama and French 1993; 2015). Therefore, these funds may not have comparable factor loadings. Different factors may underperform for long periods, and it can be psychologically challenging for an investor (see, e.g., Fama and French 2020). However, the asset pricing literature considers the market portfolio to be efficient. Therefore, investing in a broad market index may be a sensible option.

5.3 Future research opportunities

This study aimed to find if the synthetic sentiment indices constructed in the thesis are related to aggregate stock market movements. The study also examines if the Google Trends-based sentiment can explain the cross-section of asset returns. While the evidence indicates the answer for both research questions to be no, there are still paths to conduct further research. I use comprehensive regression techniques throughout the study. If the empirical findings are not significant when using these methods, adding more sophisticated correction techniques does not seem necessary. Therefore, I feel confident with the validity of the results. However, I do not use some potentially beneficial methods, such as the correction of Shanken (1992). I do not include two-pass regression methods without the rolling procedure, such as the generalized method of moments, either. Applying these methods is one avenue for further research.

The R code, which I present in Appendix 1, includes all analysis I use in the empirical part of the thesis. You can download the Google Trends data with the script I provide in the code. You can also obtain Fama–French data from the Data Library of French (2018). The study has high reliability since you can mostly replicate the results with these resources. However, The NYSE data of individual companies, collected from the Thomson Reuters Datastream, has some issues which may affect the results. For example, some companies do not have total returns or characteristics data. Also, the data may suffer from survivorship bias since Datastream does not specify if a stock delists within the period. Some of the characteristics may also suffer from look-ahead bias. Replicating the study with better firm-specific data could result in different outcomes. Issues in NYSE data also require modifying the existing regression approaches. For example, the

FMB procedure makes a deliberate compromise between the quality and amount of the regression. Better data would enable using more theoretically sound regression methods.

The generalizability leaves room for improvement since a sample of 98 keywords is far from a perfect representation of Google Trends data. On the other hand, using all available search term data may not be feasible. Different search term specifications, periods, and data frequencies offer an opportunity for future studies. The vast amount of Google Trends data offers possibilities to test different keywords and categories. Search terms with a high correlation between their search volume and the stock market index may generate more significant results. High-frequency data may also help to capture the possible short-term effects of the sentiment. Besides, using data from different periods or geographic and language regions may lead to different outcomes.

REFERENCES

Abdi, H. – Williams, L. J. (2010) Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, Vol. 2 (4), 433–459.

Aharoni, G. – Grundy, B. – Zeng, Q. (2013) Stock returns and the Miller Modigliani valuation formula: Revisiting the Fama French analysis. *Journal of Financial Economics*, Vol. 110 (2), 347–357.

Ang, A. – Hodrick, R. J. – Xing, Y. – Zhang, X. (2006) The cross-section of volatility and expected returns. *The Journal of Finance*, Vol. 61 (1), 259–299.

Antweiler, W. – Frank, M. Z. (2004) Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*, Vol. 59 (3), 1259–1294.

Baker, M. – Wurgler, J. (2006) Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, Vol. 61 (4), 1645–1680.

Baker, M. – Wurgler, J. (2007) Investor sentiment in the stock market. *Journal of Economic Perspectives*, Vol. 21 (2), 129–152.

Banz, R. W. (1981) The relationship between return and market value of common stocks. *Journal of Financial Economics*, Vol. 9 (1), 3–18.

Barber, B. M. – Odean, T. (2008) All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *The Review of Financial Studies*, Vol. 21 (2), 785–818.

Barberis, N. – Shleifer, A. – Vishny, R. (1998) A model of investor sentiment. *Journal of Financial Economics*, Vol. 49 (3), 307–343.

Basu, S. (1983) The relationship between earnings' yield, market value and return for NYSE common stocks: Further evidence. *Journal of Financial Economics*, Vol. 12 (1), 129–156.

Black, F. (1972) Capital market equilibrium with restricted borrowing. *The Journal of Business*, Vol. 45 (3), 444–455.

Black, F. – Jensen, M. C. – Scholes, M. (1972) The capital asset pricing model: Some empirical tests. In *Studies in the Theory of Capital Markets*, ed. M. C. Jensen, 79–121, Praeger Publishers Inc.

Bollen, J. – Mao, H. – Zeng, X. (2011) Twitter mood predicts the stock market. *Journal of Computational Science*, Vol. 2 (1), 1–8.

Breeden, D. T. (1979) An intertemporal asset pricing model with stochastic consumption and investment opportunities. *Journal of Financial Economics*, Vol. 7 (3), 265–296.

Brennan, M. J. – Chordia, T. – Subrahmanyam, A. (1998) Alternative factor specifications, security characteristics, and the cross-section of expected stock returns. *Journal of Financial Economics*, Vol. 49 (3), 345–373.

Brown, G. W. – Cliff, M. T. (2005) Investor sentiment and asset valuation. *The Journal of Business*, Vol. 78 (2), 405–440.

Campbell, J. Y. – Cochrane, J. H. (1999) By force of habit: A consumption-based explanation of aggregate stock market behavior. *Journal of Political Economy*, Vol. 107 (2), 205–251.

Campbell, J. Y. – Lo, A. W. – MacKinlay, A. C. (1997) *The Econometrics of Financial Markets*, Vol. 2. Princeton University Press, Princeton, NJ.

Cao, M. – Wei, J. (2005) Stock market returns: A note on temperature anomaly. *Journal of Banking & Finance*, Vol. 29 (6), 1559–1573.

Carhart, M. M. (1997) On persistence in mutual fund performance. *The Journal of Finance*, Vol. 52 (1), 57–82.

Chen, N.-F. – Roll, R. – Ross, S. A. (1986) Economic forces and the stock market. *The Journal of Business*, Vol. 59 (3), 383–403.

Chen, N.-f. – Zhang, F. (1998) Risk and return of value stocks. *The Journal of Business*, Vol. 71 (4), 501–535.

Choi, H. – Varian, H. (2012) Predicting the present with Google Trends. *Economic Record*, Vol. 88, 2–9.

Chordia, T. – Goyal, A. – Saretto, A. (2017) p-hacking: Evidence from two million trading strategies. *Swiss Finance Institute Research Paper No. 17-37*.

Cochrane, J. (2005) *Asset Pricing: Revised Edition*. Princeton University Press.

Curme, C. – Preis, T. – Stanley, H. E. – Moat, H. S. (2014) Quantifying the semantics of search behavior before stock market moves. *Proceedings of the National Academy of Sciences*, Vol. 111 (32), 11600–11605.

Da, Z. – Engelberg, J. – Gao, P. (2011) In search of attention. *The Journal of Finance*, Vol. 66 (5), 1461–1499.

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

Daniel, K. – Hirshleifer, D. – Subrahmanyam, A. (1998) Investor psychology and security market under-and overreactions. *The Journal of Finance*, Vol. 53 (6), 1839–1885.

De Bondt, W. F. – Thaler, R. (1985) Does the stock market overreact? *The Journal of Finance*, Vol. 40 (3), 793–805.

Dimpfl, T. – Jank, S. (2016) Can internet search queries help to predict stock market volatility? *European Financial Management*, Vol. 22 (2), 171–192.

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

Fama, E. F. – French, K. R. (1988) Dividend yields and expected stock returns. *Journal of Financial Economics*, Vol. 22 (1), 3–25.

Fama, E. F. – French, K. R. (1992) The cross-section of expected stock returns. *The Journal of Finance*, Vol. 47 (2), 427–465.

Fama, E. F. – French, K. R. (1993) Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, Vol. 33 (1), 3–56.

Fama, E. F. – French, K. R. (1996) Multifactor explanations of asset pricing anomalies. *The Journal of Finance*, Vol. 51 (1), 55–84.

Fama, E. F. – French, K. R. (2006) Profitability, investment and average returns. *Journal of Financial Economics*, Vol. 82 (3), 491–518.

Fama, E. F. – French, K. R. (2008) Dissecting anomalies. *The Journal of Finance*, Vol. 63 (4), 1653–1678.

Fama, E. F. – French, K. R. (2015) A five-factor asset pricing model. *Journal of Financial Economics*, Vol. 116 (1), 1–22.

Fama, E. F. – French, K. R. (2020) The value premium. Tech. rep., Fama-Miller Working Paper.

Fama, E. F. – MacBeth, J. D. (1973) Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, Vol. 81 (3), 607–636.

French, K. R. (2018) Data library. <http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html>, retrieved 29.9.2018.

Gibbons, M. R. – Ross, S. A. – Shanken, J. (1989) A test of the efficiency of a given portfolio. *Econometrica*, Vol. 57 (5), 1121–1152.

Ginsberg, J. – Mohebbi, M. H. – Patel, R. S. – Brammer, L. – Smolinski, M. S. – Brilliant, L. (2009) Detecting influenza epidemics using search engine query data. *Nature*, Vol. 457 (7232), 1012–1014.

Hirshleifer, D. – Shumway, T. (2003) Good day sunshine: Stock returns and the weather. *The Journal of Finance*, Vol. 58 (3), 1009–1032.

Hlavac, M. (2018) stargazer: Well-formatted regression and summary statistics tables. <https://CRAN.R-project.org/package=stargazer>, R package version 5.2.2, retrieved 25.10.2018.

Jegadeesh, N. – Titman, S. (1993) Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, Vol. 48 (1), 65–91.

Jegadeesh, N. – Titman, S. (2001) Profitability of momentum strategies: An evaluation of alternative explanations. *The Journal of Finance*, Vol. 56 (2), 699–720.

Lakonishok, J. – Shleifer, A. – Vishny, R. W. (1994) Contrarian investment, extrapolation, and risk. *The Journal of Finance*, Vol. 49 (5), 1541–1578.

Lintner, J. (1965) The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *The Review of Economics and Statistics*, Vol. 47 (1), 13–37.

Lo, A. W. – MacKinlay, A. C. (1990a) Data-snooping biases in tests of financial asset pricing models. *The Review of Financial Studies*, Vol. 3 (3), 431–467.

Lo, A. W. – MacKinlay, A. C. (1990b) When are contrarian profits due to stock market overreaction? *The Review of Financial Studies*, Vol. 3 (2), 175–205.

Lo, A. W. – Mamaysky, H. – Wang, J. (2000) Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. *The Journal of Finance*, Vol. 55 (4), 1705–1765.

Lucas Jr, R. E. (1978) Asset prices in an exchange economy. *Econometrica*, Vol. 46 (6), 1429–1445.

Mankiw, N. G. – Shapiro, M. D. (1986) Risk and return: Consumption beta versus market beta. *The Review of Economics and Statistics*, Vol. 68 (3), 452–459.

Markowitz, H. (1952) Portfolio selection. *The Journal of Finance*, Vol. 7 (1), 77–91.

Massicotte, P. – Eddelbuettel, D. (2018) gtrendsR: A package on performing and displaying Google Trends queries. <https://cran.r-project.org/web/packages/gtrendsR/gtrendsR.pdf>, manual for the R package gtrendsR, retrieved 27.8.2018.

Miller, M. H. – Modigliani, F. (1961) Dividend policy, growth, and the valuation of shares. *The Journal of Business*, Vol. 34 (4), 411–433.

Miller, M. H. – Scholes, M. S. (1982) Dividends and taxes: Some empirical evidence. *Journal of Political Economy*, Vol. 90 (6), 1118–1141.

Moat, H. S. – Curme, C. – Avakian, A. – Kenett, D. Y. – Stanley, H. E. – Preis, T. (2013) Quantifying Wikipedia usage patterns before stock market moves. *Scientific Reports*, Vol. 3 (1), 1–5.

Moskowitz, T. J. – Ooi, Y. H. – Pedersen, L. H. (2012) Time series momentum. *Journal of Financial Economics*, Vol. 104 (2), 228–250.

Newey, W. K. – West, K. D. (1987) A simple, positive semi-definite, heteroskedasticity and autocorrelation. *Econometrica*, Vol. 55 (3), 703–708.

Novy-Marx, R. (2013) The other side of value: The gross profitability premium. *Journal of Financial Economics*, Vol. 108 (1), 1–28.

Pastor, L. – Stambaugh, R. F. – Taylor, L. A. (2019) Sustainable investing in equilibrium. Tech. rep., National Bureau of Economic Research.

Petersen, M. A. (2009) Estimating standard errors in finance panel data sets: Comparing approaches. *The Review of Financial Studies*, Vol. 22 (1), 435–480.

Petkova, R. (2006) Do the Fama–French factors proxy for innovations in predictive variables? *The Journal of Finance*, Vol. 61 (2), 581–612.

Piotroski, J. D. (2000) Value investing: The use of historical financial statement information to separate winners from losers. *Journal of Accounting Research*, Vol. 38 (1), 1–41.

Preis, T. – Moat, H. S. – Stanley, H. E. (2013) Quantifying trading behavior in financial markets using Google Trends. *Scientific Reports*, Vol. 3 (1684), 1–6.

Preis, T. – Moat, H. S. – Stanley, H. E. – Bishop, S. R. (2012) Quantifying the advantage of looking forward. *Scientific Reports*, Vol. 2 (1), 1–2.

Preis, T. – Reith, D. – Stanley, H. E. (2010) Complex dynamics of our economic life on different scales: insights from search engine query data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol. 368 (1933), 5707–5719.

Roll, R. (1977) A critique of the asset pricing theory's tests Part I: On past and potential testability of the theory. *Journal of Financial Economics*, Vol. 4 (2), 129–176.

Rosenberg, B. – Reid, K. – Lanstein, R. (1985) Persuasive evidence of market inefficiency. *The Journal of Portfolio Management*, Vol. 11 (3), 9–16.

Ross, S. (1976) The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, Vol. 13 (3), 341–360.

Shanken, J. (1992) On the estimation of beta-pricing models. *The Review of Financial Studies*, Vol. 5 (1), 1–33.

Sharpe, W. F. (1964) Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, Vol. 19 (3), 425–442.

Sloan, R. G. (1996) Do stock prices fully reflect information in accruals and cash flows about future earnings? *The Accounting Review*, Vol. 71 (3), 289–315.

Stephens-Davidowitz, S. – Varian, H. (2014) *A Hands-on Guide to Google Data.* Google, Inc., further details on the construction can be found on the Google Trends page.

Subrahmanyam, A. (2010) The cross-section of expected stock returns: What have we learnt from the past twenty-five years of research? *European Financial Management*, Vol. 16 (1), 27–42.

Tetlock, P. C. (2007) Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, Vol. 62 (3), 1139–1168.

Thomson Reuters (2015) *Worldscope Database: Data Definitions Guide*. Thomson Reuters, 14.3 edn.

Tobin, J. (1958) Liquidity preference as behavior towards risk. *The Review of Economic Studies*, Vol. 25 (2), 65–86.

White, H. (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, Vol. 48 (4), 817–838.

Zhang, C. (2009) On the explanatory power of firm-specific variables in cross-sections of expected returns. *Journal of Empirical Finance*, Vol. 16 (2), 306–317.

APPENDICES

Appendix 1. R code

The file *main.R* includes all analysis I use in the empirical part of the thesis. Also, the code demonstrates how to generate tables and figures presented in Sections 3 and 4. The code runs successfully on a 64-bit build of R version 3.5.1. The functionality of the code may differ depending on the user's R version. The packages may also function differently depending on their version. I include all the required packages on lines 2–22. R package *stargazer* (Hlavac 2018) requires changes to its source code to show the t-statistics as I present them in tables in Section 4. I give instructions on to do this change on lines 26–23 and demonstrate how to switch between LaTeXand text-based table mode on lines 37–38. However, you do not have to perform these steps to run the code.

I highlight different sections and use comments throughout the code to make it more coherent. You should execute the code line by line to define all the required variables. I use a for-each loop in parallel when the commands require considerable computing power, which should speed up the process. Also, the console shows the progress in loops, which may take a long time to execute. However, the code runs the Fama–MacBeth regressions and long-short regressions using only the first sentiment index by default. Suppose you wish to execute code for all sentiments. In that case, you can manually change the variable and run a part of the code multiple times. I instruct this change on lines 492–497, 694, 947, and 1045.

You can download Google Trends data with the script I provide on lines 41–85. You can also obtain Fama–French data from the Data Library of French (2018). NYSE data originates from Thomson Reuters Datastream. Suppose you have access to such a database. In that case, you can follow the procedure I describe in Section 3 to retrieve the same data and replicate the results with the code. The code may not be optimal and lacks some programming leading practices, such as error handling procedures. You should, however, be able to run the code as it is.

main.R

```
#========
#PACKAGES
#========

# This section includes packages used throughout the code
# Use install.packages() if required

library(gtrendsR)
library(reshape2)
library(svMisc)
library(dplyr)
library(stats)
library(qpcR)
library(stargazer)
library(ggplot2)
library(zoo)
library(stringr)
library(DescTools)
library(multiwayvcov)
library(lmtest)
library(doParallel)
library(foreach)

# Stargazer

# Change Stargazer source code:
# Remove # from next line and select
# trace(stargazer:::.stargazer.wrap, edit = T)
# Change lines 7105 and 7016
# (Lines may differ depending on Stargazer version)
# .format.t.stats.left <- "t = "
# .format.t.stats.right <- ""
# Replace lines with
# .format.t.stats.left <- "("
# .format.t.stats.right <- ")"

# Use type = "text" for viewing tables in R
# Use type = "latex" to export table to LaTeX

#====================================
#SCRIPT TO DOWNLOAD GOOGLE TRENDS DATA
#====================================

# Google Trends search queries
queries <- c("arts", "banking", "bonds", "bubble", "buy", "cancer", "car",
             "cash", "chance", "color", "conflict", "consume", "consumption",
             "crash", "credit", "crisis", "culture", "debt", "default",
             "derivatives", "dividend", "dow jones", "earnings", "economics",
             "economy", "energy", "environment", "fed", "finance",
             "financial markets", "fine", "fond", "food", "forex", "freedom",
             "fun", "gain", "gains", "garden", "gold", "greed", "growth",
             "happy", "headlines", "health", "hedge", "holiday", "home",
             "house", "housing", "inflation", "invest", "investment",
             "kitchen", "labor", "leverage", "lifestyle", "loss", "markets",
             "marriage", "metals", "money", "movie", "nasdaq", "nyse", "office",
```

```
56              "oil", "opportunity", "ore", "politics", "portfolio", "present",
57              "profit", "rare earths", "religion", "restaurant", "return",
58              "returns", "revenue", "rich", "ring", "risk", "sell", "short sell",
59              "short selling", "society", "stock market", "stocks", "success",
60              "tourism", "trader", "train", "transaction", "travel",
61              "unemployment", "war", "water", "world")
62
63 # Custom date format
64 period <- "2004-01-01 2017-12-31"
65
66 # Download Search Volume Index for each search query individually
67 SVI = c()
68 for(i in 1:length(queries)){
69
70   tmp <- gtrends(queries[i],
71                  geo = "US",
72                  time =  period)$interest_over_time[, c(1,3,2)]
73   SVI <- rbind(SVI, tmp)
74
75   progress(i, max.value = length(queries))
76   if (i == length(queries)) message("Done")
77   i = i + 1
78 }
79
80 SVI <- dcast(SVI, date ~ keyword, value.var = "hits")
81 SVI$date <- as.Date(SVI$date)
82
83 # Write a CSV file
84 csvFileName <- paste("SVI_", Sys.Date(), ".csv", sep = "")
85 write.csv2(SVI, file = csvFileName, row.names = FALSE)
86
87 #=====================
88 #KENNETH FRENCH'S DATA
89 #=====================
90
91 # 25 portfolios from Kenneth French's Data Library
92 portfolios <- read.csv("~/R/25_Portfolios_5x5.csv", sep = ",", header = TRUE)
93
94 # Five factors from Kenneth French's Data Library
95 rs_factors <- read.csv("~/R/F-F_Research_Data_5_Factors_2x3.csv",
96                        sep = ",", header = TRUE)
97
98 # Select a time period
99 portfolios <- portfolios %>% filter(between(Date, 200401, 201712))
100 rs_factors <- rs_factors %>% filter(between(Date, 200401, 201712))
101
102 portfolios_backup <- portfolios
103
104 # Assign dates for further use
105 date_list <- portfolios$Date
106
107 #=================
108 #SENTIMENT INDICES
109 #=================
110
111 # Loads a file averaged over 15 unique realizations of data
112 # Google Data may change slightly daily
```

```
# One may alternatively use a single realization downloaded with a provided script
SVI <- read.csv2("~/R/SVI.csv")

# Replace spaces with dots to avoid issues with data
names(SVI) <- gsub(" ", ".", names(SVI))

# SVI samples
# Query "rare earths" is not included since it has a volume of zero at times
# I sort queries based on their performance according to Preis, Moat and Stanley
    (2013)

# Whole sample of search queries
SVI_sample1 <- dplyr::select(SVI, debt, color, stocks, restaurant, portfolio,
                             inflation, housing, dow.jones, revenue, economics,
                             credit, markets, return, unemployment, money,
                             religion, cancer, growth, investment, hedge,
                             marriage, bonds, derivatives, headlines, profit,
                             society, leverage, loss, cash, office, fine,
                             stock.market, banking, crisis, happy, car, nasdaq,
                             gains, finance, sell, invest, fed, house, metals,
                             travel, returns, gain, default, present, holiday,
                             water, rich, risk, gold, success, oil, war,
                             economy, chance, short.sell, lifestyle, greed,
                             food, financial.markets, movie, nyse, ore,
                             opportunity, health, short.selling, earnings, arts,
                             culture, bubble, buy, trader, tourism, politics,
                             energy, consume, consumption, freedom, dividend,
                             world, conflict, kitchen, forex, home, crash,
                             transaction, garden, fond, train, labor, fun,
                             environment, ring)

# 15 best performing search queries
SVI_sample2 <- dplyr::select(SVI, debt, color, stocks, restaurant, portfolio,
                             inflation, housing, dow.jones, revenue, economics,
                             credit, markets, return, unemployment, money)

# 15 search queries with the highest relative keyword occurrence
SVI_sample3 <- dplyr::select(SVI, hedge, dividend, earnings, inflation, markets,
                             bonds, debt, financial.markets, gains, investment,
                             growth, derivatives, crisis, unemployment, banking)

# Best performing search query (debt)
SVI_sample4 <- dplyr::select(SVI, debt)

SVI_list <- list(SVI_sample1, SVI_sample2, SVI_sample3, SVI_sample4)

# Sentiment indices
S_list = c()
PC_list = c()
for (i in 1:length(SVI_list)) {
  SVI_sample <- SVI_list[[i]]

  SVI_matrix <- matrix(as.numeric(unlist(SVI_sample[, 1:ncol(SVI_sample)])),
                       nrow = nrow(SVI_sample))
  SVI_matrix <- diff(log(SVI_matrix))
  PC <- princomp(SVI_matrix, cor = FALSE, scores = TRUE)
  S <- SVI_matrix %*% as.matrix(PC$loadings[, 1])
```

```
  S <- round(S, 4)
  S <- c(NA, S)

  S_list <- cbind(S_list, S)
  PC_list <- qpcR:::cbind.na(PC_list, PC$sdev^2)
}

S_list <- as.data.frame(cbind(date_list, S_list))
colnames(S_list) <- c("Date", "S1", "S2", "S3", "S4")

# Sentiment indices with lag
S_list_lag1 <- head(cbind(c(NA, S_list$S1), c(NA, S_list$S2),
                          c(NA, S_list$S3),c(NA, S_list$S4)), 168)

S_list_lag1 <- as.data.frame(cbind(date_list, S_list_lag1))
colnames(S_list_lag1) <- c("Date", "S1_lag1", "S2_lag1", "S3_lag1", "S4_lag1")

# Principal components
PC_list <- PC_list[, 2:4]
PC_list <- as.data.frame(cbind(as.data.frame(seq(1,97,1)), PC_list))
colnames(PC_list) <- c("Component", "Sentiment 1", "Sentiment 2", "Sentiment 3")
PC_list <- melt(PC_list, id = "Component")
PC_list <- na.omit(PC_list)

#=========
#NYSE DATA
#=========

# This part of the code shows how to construct the NYSE_panel.csv
# Data is from Thomson Reuters Datastream
# Data files do not include firms with missing data
# Data files do not include firms whose SIC code begins with the number six

# Data
#RI <- read.csv2("~/R/NYSE_data/RI.csv")
#MV <- read.csv2("~/R/NYSE_data/MV.csv")
#PTBV <- read.csv2("~/R/NYSE_data/PTBV.csv")
#PE <- read.csv2("~/R/NYSE_data/PE.csv")
#PC <- read.csv2("~/R/NYSE_data/PC.csv")
#DY <- read.csv2("~/R/NYSE_data/DY.csv")

# Panel format
#RI <- melt(RI, id.vars = "Date")
#MV <- melt(MV, id.vars = "Date")
#PTBV <- melt(PTBV, id.vars = "Date")
#PE <- melt(PE, id.vars = "Date")
#PC <- melt(PC, id.vars = "Date")
#DY <- melt(DY, id.vars = "Date")

#NYSE_panel <- cbind(RI, MV$value, PTBV$value, PE$value, PC$value, DY$value)
#colnames(NYSE_panel) <- c("Date", "variable", "RI", "MV", "PB", "PE", "PC", "DY")

# Write a CSV file
#write.csv2(NYSE_panel, "NYSE_panel.csv", row.names = FALSE)

#============
#NYSE FACTORS
```

```
#============

# NYSE data
NYSE_panel <- read.csv2("~/R/NYSE_panel.csv")

uniqid <- unique(NYSE_panel$Date) %>% sort() %>% data.frame("Date" = .) %>%
  mutate(id = seq(1, length(unique(NYSE_panel$Date)), 1))
NYSE_panel <- merge(NYSE_panel, uniqid, by = "Date")

# Compile NYSE factors
tmp = c()
ls_factors = c()
for (j in 1:5) {
  for(i in min(NYSE_panel$id):max(NYSE_panel$id)){
    tmp_data <- na.omit(NYSE_panel)
    tmp_data <- filter(tmp_data, id == i)
    characteristics <- list(tmp_data$MV, tmp_data$PB, tmp_data$PE,
                            tmp_data$PC, tmp_data$DY)
    characteristic <- unlist(characteristics[j])
    if (j < 5) {
      tmp_data <- tmp_data %>% mutate(decile = ntile(characteristic, 10))
      A <- tmp_data %>% filter(decile < 4)
      B <- tmp_data %>% filter(decile > 7)
    } else {
      A <- tmp_data %>% filter(DY > 0)
      B <- tmp_data %>% filter(DY == 0)
    }
    difference <- round(mean(A$RI) - mean(B$RI), 4)
    tmp <- rbind(tmp, difference)
  }
  tmp <- as.vector(tmp)
  ls_factors <- cbind(ls_factors, tmp)
  tmp = c()
  progress(j, max.value = 5)
  if (j == 5) message("Done")
  j = j + 1
}

NYSE_panel$id <- NULL

ls_factors <- as.data.frame(ls_factors)
ls_factors <- cbind(date_list, ls_factors)
colnames(ls_factors) <- c("Date", paste("F_", names(NYSE_panel[4:8]), sep = ""))

# Factor names
factornames <- unique(c(names(rs_factors[2:6]), names(S_list[2:5])))

#============================
#DATA DESCRIPTIVE STATISTICS
#============================

# Portfolio statistics
stargazer(round(matrix(colMeans(portfolios[2:ncol(portfolios)]),
                       nrow = 5, ncol = 5), 3), header = FALSE, type = "latex")
stargazer(round(matrix(sqrt(var(portfolios[2:ncol(portfolios)])),
                       nrow = 5, ncol = 5), 3), header = FALSE, type = "latex")

```

```
# Other statistics
# NYSE data
NYSE <- read.csv2("~/R/NYSE_panel.csv")
stargazer(NYSE[2:ncol(NYSE)],
          header = FALSE,
          summary = TRUE,
          no.space = TRUE,
          type = "latex")

# NYSE factors
stargazer(ls_factors[2:ncol(ls_factors)],
          header = FALSE,
          summary = TRUE,
          no.space = TRUE,
          type = "latex")

# Fama & French five factors
stargazer(rs_factors[2:ncol(rs_factors)],
          header = FALSE,
          summary = TRUE,
          no.space = TRUE,
          type = "latex")

# Sentiment and lagged sentiment
stargazer(cbind(S_list[2:ncol(S_list)], S_list_lag1[2:ncol(S_list_lag1)]),
          header = FALSE,
          summary = TRUE,
          no.space = TRUE,
          type = "latex")

# Correlation matrix
stargazer(cor(ls_factors[2:ncol(ls_factors)]),
          header = FALSE,
          no.space = TRUE,
          type = "latex")

stargazer(cor(rs_factors[2:6]),
          header = FALSE,
          no.space = TRUE,
          type = "latex")

stargazer(cor(na.omit(S_list[2:5])),
          header = FALSE,
          no.space = TRUE,
          type = "latex")

# Plot market return
mkt_ret <- rs_factors$Mkt.RF + rs_factors$RF
cum_mkt_ret <- as.numeric(round(sapply(as.data.frame(mkt_ret) / 100,
                                       function(x) cumprod(1 + x) - 1) * 100, 4))
market <- merge(cbind(date_list, mkt_ret),
                cbind(date_list, cum_mkt_ret), id = "Date")
colnames(market) <- c("Time", "Monthly market return", "Cumulative market return")
market$Time <- as.yearmon(paste0(str_sub(market$Time, 1, 4), "-",
                                 str_sub(market$Time, -2, -1)))
market <- melt(market, id = "Time")
```

```
ggplot(market) + geom_line(aes(x = Time, y = value)) + facet_wrap( ~ variable,
    scales = "free") + ylab("Return") + theme_bw(base_size = 20)
ggsave("Market.pdf", width = 30, height = 20, units = "cm")

# Plot Search Volume Indices
SVI_tmp <- read.csv2("~/R/SVI.csv")
colnames(SVI_tmp) <- gsub("\\.", " ", colnames(SVI_tmp))
SVI_tmp <- melt(SVI_tmp, id = "date")
ggplot(SVI_tmp) + geom_boxplot(aes(x = reorder(variable, value, FUN = median), y =
     value)) + geom_hline(yintercept = mean(SVI_tmp$value), linetype="dashed") +
    xlab("Search query") + ylab("Search Volume Index") + coord_flip() + theme_bw(
    base_size = 12)
ggsave("SVI.pdf", width = 20, height = 30, units = "cm")

# Plot Search Volume Index examples
SVI_example <- SVI_tmp %>% filter(variable %in% c("debt", "hedge", "holiday",
                                                  "invest", "nyse",
                                                  "short selling")) %>%
  arrange(match(variable, c("debt", "hedge", "holiday", "invest",
                            "nyse", "short selling")), desc(date), desc(value))
colnames(SVI_example)[1] <- "Time"
SVI_example$Time <- as.yearmon(SVI_example$Time)
ggplot(SVI_example) + geom_line(aes(x = Time, y = value)) + facet_wrap( ~ variable
    , scales="free") + ylab("Search Volume Index") + theme_bw(base_size = 20)
ggsave("SVI_example.pdf", width = 30, height = 20, units = "cm")

# Plot principal components
ggplot(PC_list) + geom_col(aes(x = Component, y = value)) + facet_wrap( ~ variable
    , scales="free") + ylab("Variance") + theme_bw(base_size = 20)
ggsave("PC_variance.pdf", width = 30, height = 20, units = "cm")

# Plot sentiments
S_list_tmp <- S_list
colnames(S_list_tmp) <- c("Time", "Sentiment 1", "Sentiment 2",
                          "Sentiment 3", "Sentiment 4")
S_list_tmp$Time <- as.yearmon(paste0(str_sub(S_list_tmp$Time, 1, 4), "-",
                                     str_sub(S_list_tmp$Time, -2, -1)))
ggplot(melt(S_list_tmp, id = "Time")) + geom_line(aes(x = Time, y = value)) +
    facet_wrap( ~ variable, scales="free") + ylab("Value") + theme_bw(base_size =
    20)
ggsave("Sentiments.pdf", width = 30, height = 20, units = "cm")

# Plot Fama & French 25 portfolios
portfolios_tmp <- melt(portfolios, id = "Date")
portfolios_tmp$variable <- gsub("\\.", " ", portfolios_tmp$variable)
ggplot(portfolios_tmp) + geom_boxplot(aes(x = variable, y = value)) + xlab("
    Portfolio") + ylab("Return") + theme_bw(base_size = 20) + theme(axis.text.x =
    element_text(angle = 90, hjust = 1))
ggsave("FF25_boxplot.pdf", width = 30, height = 20, units = "cm")

#================================
#NYSE REGRESSIONS WITH CLUSTERING
#================================

# Combine data
NYSE_panel <- merge(NYSE_panel, rs_factors, by = "Date")

```

```
# Excess return
NYSE_panel$RI.RF <- NYSE_panel$RI - NYSE_panel$RF

NYSE_panel <- merge(NYSE_panel, ls_factors, by = "Date")
NYSE_panel <- merge(NYSE_panel, S_list, by = "Date" )

# Omit NAs
NYSE_panel <- na.omit(NYSE_panel)
rownames(NYSE_panel) <- NULL

# Add unique firm and month indices
uniqid <- unique(NYSE_panel$variable) %>%
  sort() %>% data.frame("variable" = .) %>%
  mutate(firmid = seq(1, length(unique(NYSE_panel$variable)), 1))

NYSE_panel <- merge(NYSE_panel, uniqid, by = "variable")

uniqid <- unique(NYSE_panel$Date) %>%
  sort() %>% data.frame("Date" = .) %>%
  mutate(timeid = seq(1, length(unique(NYSE_panel$Date)), 1))

NYSE_panel <- merge(NYSE_panel, uniqid, by = "Date")

# Arrange data
NYSE_panel <- arrange(NYSE_panel, firmid, timeid)

# Regressions
OLS1 <- lm(RI.RF ~ S1 + F_MV + F_PB + F_PE + F_PC + F_DY, data = NYSE_panel)
OLS2 <- lm(RI.RF ~ S2 + F_MV + F_PB + F_PE + F_PC + F_DY, data = NYSE_panel)
OLS3 <- lm(RI.RF ~ S3 + F_MV + F_PB + F_PE + F_PC + F_DY, data = NYSE_panel)
OLS4 <- lm(RI.RF ~ S4 + F_MV + F_PB + F_PE + F_PC + F_DY, data = NYSE_panel)

# Regression statistics
covariate_labels <- list("Constant", "Sentiment 1", "Sentiment 2",
                         "Sentiment 3", "Sentiment 4", "MV SMB",
                         "PB Low-High", "PE Low-High", "PC Low-High", "DY >0-=0")

N <- format(nrow(NYSE_panel), big.mark = ",")

R2a1 <- round(summary(OLS1)$adj.r.squared, 3)
R2a2 <- round(summary(OLS2)$adj.r.squared, 3)
R2a3 <- round(summary(OLS3)$adj.r.squared, 3)
R2a4 <- round(summary(OLS4)$adj.r.squared, 3)

# Clustering
Panel_A <- list(
  coeftest(OLS1, cluster.vcov(OLS1, ~ firmid)),
  coeftest(OLS2, cluster.vcov(OLS2, ~ firmid)),
  coeftest(OLS3, cluster.vcov(OLS3, ~ firmid)),
  coeftest(OLS4, cluster.vcov(OLS4, ~ firmid)),
  coeftest(OLS1, cluster.vcov(OLS1, ~ timeid)),
  coeftest(OLS2, cluster.vcov(OLS2, ~ timeid)))

Panel_B <- list(
  coeftest(OLS3, cluster.vcov(OLS3, ~ timeid)),
  coeftest(OLS4, cluster.vcov(OLS4, ~ timeid)),
  coeftest(OLS1, cluster.vcov(OLS1, ~ firmid + timeid)),
```

```
  coeftest(OLS2, cluster.vcov(OLS2, ~ firmid + timeid)),
  coeftest(OLS3, cluster.vcov(OLS3, ~ firmid + timeid)),
  coeftest(OLS4, cluster.vcov(OLS4, ~ firmid + timeid)))

# Create a table
# Panel A
stargazer(Panel_A,
          header = FALSE,
          intercept.bottom = FALSE,
          model.names = FALSE,
          model.numbers = FALSE,
          dep.var.labels.include = FALSE,
          no.space = TRUE,
          title = "Regressions",
          dep.var.caption  = "NYSE panel: Regressions I--IV",
          column.labels = c("(I)", "(II)", "(III)", "(IV)", "(V)", "(VI)"),
          covariate.labels = c(unlist(covariate_labels)),
          add.lines = list(
            c("N", N, N, N, N, N, N),
            c("Robust std. errors", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes"),
            c("Clustering", "Firm", "Firm", "Firm", "Firm", "Month", "Month"),
            c("Adjusted R2", R2a1, R2a2, R2a3, R2a4, R2a1, R2a2)),
          report=('vc*t'),
          type = "latex")

# Panel B
stargazer(Panel_B,
          header = FALSE,
          intercept.bottom = FALSE,
          model.names = FALSE,
          model.numbers = FALSE,
          dep.var.labels.include = FALSE,
          no.space = TRUE,
          title = "Regressions",
          dep.var.caption  = "NYSE panel: Regressions I--IV",
          column.labels = c("(VII)", "(VIII)", "(IX)", "(X)", "(XI)", "(XII)"),
          covariate.labels = c(unlist(covariate_labels)),
          add.lines = list(
            c("N", N, N, N, N, N, N),
            c("Robust std. errors", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes"),
            c("Clustering", "Month", "Month", "F/M", "F/M", "F/M", "F/M"),
            c("Adjusted R2", R2a3, R2a4, R2a1, R2a2, R2a3, R2a4)),
          report=('vc*t'),
          type = "latex")

#======================
#LONG-SHORT REGRESSIONS
#======================

# i = 1 runs code for the first sentiment
# Change i to run code for the other three sentiments
i = 1

S <- list(NYSE_panel$S1, NYSE_panel$S2, NYSE_panel$S3, NYSE_panel$S4)

N <- format(nrow(NYSE_panel), big.mark = ",")

```

```
# Long-Short with sentiment only
LS1.1 <- lm(RI.RF ~ unlist(S[i]), data = NYSE_panel)
LS1.2 <- lm(F_MV ~ unlist(S[i]), data = NYSE_panel)
LS1.3 <- lm(F_PB ~ unlist(S[i]), data = NYSE_panel)
LS1.4 <- lm(F_PE ~ unlist(S[i]), data = NYSE_panel)
LS1.5 <- lm(F_PC ~ unlist(S[i]), data = NYSE_panel)
LS1.6 <- lm(F_DY ~ unlist(S[i]), data = NYSE_panel)

Double_cluster1 <- list(
  coeftest(LS1.1, cluster.vcov(LS1.1, ~ firmid + timeid)),
  coeftest(LS1.2, cluster.vcov(LS1.2, ~ firmid + timeid)),
  coeftest(LS1.3, cluster.vcov(LS1.3, ~ firmid + timeid)),
  coeftest(LS1.4, cluster.vcov(LS1.4, ~ firmid + timeid)),
  coeftest(LS1.5, cluster.vcov(LS1.5, ~ firmid + timeid)),
  coeftest(LS1.6, cluster.vcov(LS1.6, ~ firmid + timeid)))

R2_adjusted1 <- list(
  round(summary(LS1.1)$adj.r.squared, 3),
  round(summary(LS1.2)$adj.r.squared, 3),
  round(summary(LS1.3)$adj.r.squared, 3),
  round(summary(LS1.4)$adj.r.squared, 3),
  round(summary(LS1.5)$adj.r.squared, 3),
  round(summary(LS1.6)$adj.r.squared, 3))

# Create a table
stargazer(Double_cluster1,
          header = FALSE,
          intercept.bottom = FALSE,
          model.names = FALSE,
          model.numbers = FALSE,
          dep.var.labels.include = FALSE,
          no.space = TRUE,
          title = "Long-Short Regressions",
          dep.var.caption  = "NYSE panel: Long-Short Regressions I--VI",
          column.labels = c("(I)", "(II)", "(III)", "(IV)", "(V)", "(VI)"),
          covariate.labels = c("Constant", "Sentiment 1"),
          add.lines = list(
            c("N", N, N, N, N, N, N),
            c("Robust std. errors", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes"),
            c("Clustering", "F/M", "F/M", "F/M", "F/M", "F/M", "F/M"),
            c("Adjusted R2", unlist(R2_adjusted1))),
          report=('vc*t'),
          type = "latex")

# Long-Short with FF5 factors
LS2.1 <- lm(RI.RF ~ unlist(S[i]) + Mkt.RF + SMB + HML + RMW + CMA,
            data = NYSE_panel)
LS2.2 <- lm(F_MV ~ unlist(S[i]) + Mkt.RF + HML + RMW + CMA,
            data = NYSE_panel)
LS2.3 <- lm(F_PB ~ unlist(S[i]) + Mkt.RF + SMB + RMW + CMA,
            data = NYSE_panel)
LS2.4 <- lm(F_PE ~ unlist(S[i]) + Mkt.RF + SMB + HML + RMW + CMA,
            data = NYSE_panel)
LS2.5 <- lm(F_PC ~ unlist(S[i]) + Mkt.RF + SMB + HML + RMW + CMA,
            data = NYSE_panel)
LS2.6 <- lm(F_DY ~ unlist(S[i]) + Mkt.RF + SMB + HML + RMW + CMA,
            data = NYSE_panel)
```

```
557
558 Double_cluster2 <- list(
559   coeftest(LS2.1, cluster.vcov(LS2.1, ~ firmid + timeid)),
560   coeftest(LS2.2, cluster.vcov(LS2.2, ~ firmid + timeid)),
561   coeftest(LS2.3, cluster.vcov(LS2.3, ~ firmid + timeid)),
562   coeftest(LS2.4, cluster.vcov(LS2.4, ~ firmid + timeid)),
563   coeftest(LS2.5, cluster.vcov(LS2.5, ~ firmid + timeid)),
564   coeftest(LS2.6, cluster.vcov(LS2.6, ~ firmid + timeid)))
565
566 R2_adjusted2 <- list(
567   round(summary(LS2.1)$adj.r.squared, 3),
568   round(summary(LS2.2)$adj.r.squared, 3),
569   round(summary(LS2.3)$adj.r.squared, 3),
570   round(summary(LS2.4)$adj.r.squared, 3),
571   round(summary(LS2.5)$adj.r.squared, 3),
572   round(summary(LS2.6)$adj.r.squared, 3))
573
574 # Create a table
575 stargazer(Double_cluster2,
576           header = FALSE,
577           intercept.bottom = FALSE,
578           model.names = FALSE,
579           model.numbers = FALSE,
580           dep.var.labels.include = FALSE,
581           single.row = TRUE,
582           title = "Long-Short Regressions",
583           dep.var.caption  = "NYSE panel: Long-Short Regressions I--VI",
584           column.labels = c("(I)", "(II)", "(III)", "(IV)", "(V)", "(VI)"),
585           covariate.labels = c("Constant", "Sentiment 1", "RM-RF",
586                                "SMB", "HML", "RMW", "CMA"),
587           add.lines = list(
588             c("N", N, N, N, N, N, N),
589             c("Robust std. errors", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes"),
590             c("Clustering", "F/M", "F/M", "F/M", "F/M", "F/M", "F/M"),
591             c("Adjusted R2", unlist(R2_adjusted2))),
592           report=('vc*t'),
593           type = "latex")
594
595 #============================================================
596 #FAMA & MACBETH REGRESSIONS (NYSE STOCKS) WITHOUT SENTIMENT
597 #============================================================
598
599 NYSE_panel_backup <- NYSE_panel
600
601 # 1st stage
602 minwind = 60
603
604 cl <- makeCluster(detectCores(all.tests = TRUE, logical = FALSE))
605 registerDoParallel(cl)
606
607 coef = c()
608 betanames <- paste("B_", c(factornames[1:5]), sep = "")
609 k = 1
610 for(i in unique(NYSE_panel$firmid)){
611   tmp <- filter(NYSE_panel, firmid == i)
612   coef_tmp <- foreach(j = (minwind + 1):max(NYSE_panel$timeid), .combine = rbind)
        %dopar% {
```

```
613     regtemp <- tmp[tmp $timeid >= j - minwind & tmp$timeid <= j - 1, ]
614     if(sum(table(regtemp$RI.RF)) >= 45){
615       regtemp <- as.data.frame(regtemp)
616       coef_tmp <- lm(RI.RF ~ Mkt.RF + SMB + HML + RMW + CMA,
617                      data = regtemp, na.action = na.exclude)$coefficients
618       coef_tmp <- round(coef_tmp, 4)
619       coef_tmp <- data.frame(t(coef_tmp[-1]), j, i)
620     }
621   }
622   coef <- rbind(coef, coef_tmp)
623   if(k == length(unique(NYSE_panel$firmid))){
624     coef <- as.data.frame(coef)
625     colnames(coef) <- c(betanames, "timeid", "firmid")
626   }
627   coef_tmp = c()
628   progress(k, max.value = length(unique(NYSE_panel$firmid)))
629   if (i == length(unique(NYSE_panel$firmid))) message("Done")
630   k = k + 1
631 }
632
633 # Merging
634 NYSE_panel <- merge(NYSE_panel, coef, by = c("timeid", "firmid"))
635 NYSE_panel <- NYSE_panel %>% arrange(firmid, timeid)
636
637 # 2nd stage
638 lambdas = c()
639 for(i in unique(NYSE_panel$timeid)){
640   fit = c()
641   tmp <- NYSE_panel %>% filter(timeid == i)
642   fit <- lm(RI.RF ~ B_Mkt.RF + B_SMB + B_HML + B_RMW + B_CMA, data = tmp)
643   lambdas <- rbind(lambdas, cbind(t(fit$coefficients)))
644   progress(i, max.value = length(unique(NYSE_panel$timeid)))
645   if (i == length(unique(NYSE_panel$timeid))) message("Done")
646   i = i + 1
647 }
648
649 # Statistics
650 reg_tmp <- lm(RI.RF ~ B_Mkt.RF + B_SMB + B_HML + B_RMW + B_CMA,
651               data = NYSE_panel)
652 N <- format(nrow(NYSE_panel), big.mark = ",")
653 R2 <- round(summary(reg_tmp)$r.squared, 3)
654 R2_adjusted <- round(summary(reg_tmp)$adj.r.squared, 3)
655
656 tmp3 = c()
657 for(i in 1:6){
658   coefs <- round(mean(lambdas[, i]), 3)
659   t.values <- round((mean(lambdas[, i])) / (sd(lambdas[, i]) / sqrt(length(lambdas
        [, i])))), 3)
660   t.values <- paste0("(", t.values,")")
661   tmp3 <- rbind(tmp3, coefs, t.values)
662 }
663
664 rownames(tmp3) <- c("Constant", "t ", "lambda_RM-RF", "t",
665                     "lambda_SMB","t", "lambda_HML", "t",
666                     "lambda_RMW", "t", "lambda_CMA", "t")
667
668 # Create a table
```

```
669 stargazer(tmp3,
670           header = FALSE,
671           rownames = TRUE,
672           colnames = TRUE,
673           type = "latex")
674
675 #=======================================
676 #FAMA & MACBETH REGRESSIONS (NYSE STOCKS)
677 #=======================================
678
679 NYSE_panel <- NYSE_panel_backup
680
681 # 1st stage
682 minwind = 60
683
684 cl <- makeCluster(detectCores(all.tests = TRUE, logical = FALSE))
685 registerDoParallel(cl)
686
687 coef = c()
688 betanames <- paste("B_", c(factornames[1:5], "S"), sep = "")
689 k = 1
690 for(i in unique(NYSE_panel$firmid)){
691   tmp <- filter(NYSE_panel, firmid == i)
692   coef_tmp <- foreach(j = (minwind + 1):max(NYSE_panel$timeid), .combine = rbind)
        %dopar% {
693     regtemp <- tmp[tmp $timeid >= j - minwind & tmp$timeid <= j - 1, ]
694     S <- regtemp$S1 # Change (S1, S2, S3, S4)
695     if(sum(table(regtemp$RI.RF)) >= 45){
696       regtemp <- as.data.frame(regtemp)
697       coef_tmp <- lm(RI.RF ~ S + Mkt.RF + SMB + HML + RMW + CMA,
698                      data = regtemp, na.action = na.exclude)$coefficients
699       coef_tmp <- round(coef_tmp, 4)
700       coef_tmp <- data.frame(t(coef_tmp[-1]), j, i)
701     }
702   }
703   coef <- rbind(coef, coef_tmp)
704   if(k == length(unique(NYSE_panel$firmid))){
705     coef <- as.data.frame(coef)
706     colnames(coef) <- c(betanames, "timeid", "firmid")
707   }
708   coef_tmp = c()
709   progress(k, max.value = length(unique(NYSE_panel$firmid)))
710   if (i == length(unique(NYSE_panel$firmid))) message("Done")
711   k = k + 1
712 }
713
714 # Merging
715 NYSE_panel <- merge(NYSE_panel, coef, by = c("timeid", "firmid"))
716 NYSE_panel <- NYSE_panel %>% arrange(firmid, timeid)
717
718 # 2nd stage
719 lambdas = c()
720 for(i in unique(NYSE_panel$timeid)){
721   fit = c()
722   tmp <- NYSE_panel %>% filter(timeid == i)
723   fit <- lm(RI.RF ~ B_S + B_Mkt.RF + B_SMB + B_HML + B_RMW + B_CMA, data = tmp)
724   lambdas <- rbind(lambdas, cbind(t(fit$coefficients)))
```

```
  progress(i, max.value = length(unique(NYSE_panel$timeid)))
  if (i == length(unique(NYSE_panel$timeid))) message("Done")
  i = i + 1
}

# Statistics
reg_tmp <- lm(RI.RF ~ B_S + B_Mkt.RF + B_SMB + B_HML + B_RMW + B_CMA,
              data = NYSE_panel)
N <- format(nrow(NYSE_panel), big.mark = ",")
R2 <- round(summary(reg_tmp)$r.squared, 3)
R2_adjusted <- round(summary(reg_tmp)$adj.r.squared, 3)

tmp3 = c()
for(i in 1:7){
  coefs <- round(mean(lambdas[, i]), 3)
  t.values <- round((mean(lambdas[, i])) / (sd(lambdas[, i]) / sqrt(length(lambdas
      [, i])))), 3)
  t.values <- paste0("(", t.values,")")
  tmp3 <- rbind(tmp3, coefs, t.values)
}

rownames(tmp3) <- c("Constant", "t ", "lambda_S", "t", "lambda_RM-RF", "t",
                    "lambda_SMB","t", "lambda_HML", "t",
                    "lambda_RMW", "t", "lambda_CMA", "t")

# Create a table
stargazer(tmp3,
          header = FALSE,
          rownames = TRUE,
          colnames = TRUE,
          type = "latex")

#=======================================
#RISK-ADJUSTED REGRESSIONS (NYSE STOCKS)
#=======================================

# Risk-adjusted return
NYSE_panel$RAR <- NYSE_panel$RI.RF -
  NYSE_panel$Mkt.RF * NYSE_panel$B_Mkt.RF -
  NYSE_panel$SMB * NYSE_panel$B_SMB -
  NYSE_panel$HML * NYSE_panel$B_HML -
  NYSE_panel$RMW * NYSE_panel$B_RMW -
  NYSE_panel$CMA * NYSE_panel$B_CMA


# 2nd stage
lambdas = c()
for(i in unique(NYSE_panel$timeid)){
  fit = c()
  tmp <- NYSE_panel %>% filter(timeid == i)
  fit <- lm(RAR ~ B_S + MV + PB + PE + PC + DY, data = tmp)
  lambdas <- rbind(lambdas, cbind(t(fit$coefficients)))
  progress(i, max.value = length(unique(NYSE_panel$timeid)))
  if (i == length(unique(NYSE_panel$timeid))) message("Done")
  i = i + 1
}

```

```
# Statistics
reg_tmp <- lm(RAR ~ B_S + MV + PB + PE + PC + DY, data = NYSE_panel)
N <- format(nrow(NYSE_panel), big.mark = ",")
R2 <- round(summary(reg_tmp)$r.squared, 3)
R2_adjusted <- round(summary(reg_tmp)$adj.r.squared, 3)

tmp3 = c()
for(i in 1:7){
  coefs <- round(mean(lambdas[, i]), 3)
  t.values <- round((mean(lambdas[, i])) / (sd(lambdas[, i]) / sqrt(length(lambdas
      [, i]))), 3)
  t.values <- paste0("(", t.values,")")
  tmp3 <- rbind(tmp3, coefs, t.values)
}

rownames(tmp3) <- c("Constant", "t", "lambda_S", "t", "B_MV", "t", "B_PB",
                    "t", "B_PE", "t", "B_PC", "t", "B_DY", "t")
# Create a table
stargazer(tmp3,
          header = FALSE,
          rownames = TRUE,
          colnames = TRUE,
          type = "latex")

#=============================================================================
#FAMA & MACBETH REGRESSIONS (FAMA & FRENCH 25 PORTFOLIOS) WITHOUT SENTIMENT
#=============================================================================

# Panel format
paneldata <- melt(portfolios_backup, id.vars = "Date")

# Add Fama & French five factors and excess return
paneldata <- merge(paneldata, rs_factors, by = "Date")
paneldata$PF.RF <- paneldata$value - paneldata$RF

# Omit NAs
paneldata <- na.omit(paneldata)
rownames(paneldata) <- NULL

# Add unique firm and month indices
uniqid <- unique(paneldata$variable) %>% sort() %>% data.frame("variable" = .) %>%
  mutate(firmid = seq(1, length(unique(paneldata$variable)), 1))

paneldata <- merge(paneldata, uniqid, by = "variable")

uniqid <- unique(paneldata$Date) %>% sort() %>% data.frame("Date" = .) %>%
  mutate(timeid = seq(1, length(unique(paneldata$Date)), 1))

paneldata <- merge(paneldata, uniqid, by = "Date")

# Arrange data
paneldata <- arrange(paneldata, firmid, timeid)

# 1st stage
minwind = 60

cl <- makeCluster(detectCores(all.tests = TRUE, logical = FALSE))
```

```
registerDoParallel(cl)

coef = c()
betanames = paste("B_", factornames[1:5], sep = "")
k = 1
for(i in unique(paneldata$firmid)){
  tmp <- filter(paneldata, firmid == i)
  coef_tmp <- foreach(j = (minwind + 1):max(paneldata$timeid), .combine = rbind) %dopar% {
    regtemp <- tmp[tmp$timeid >= j - minwind & tmp$timeid <= j - 1, ]
    coef_tmp <- lm(PF.RF ~ Mkt.RF + SMB + HML + RMW + CMA, data = regtemp)$coefficients
    coef_tmp <- data.frame(t(coef_tmp[-1]), j, i)
  }
  coef <- rbind(coef, coef_tmp)
  if(k == length(unique(paneldata$firmid))){
    coef = as.data.frame(coef)
    colnames(coef) = c(betanames, "timeid", "firmid")
  }
  coef_tmp = c()
  progress(k, max.value = length(unique(paneldata$firmid)))
  if (i == length(unique(paneldata$firmid))) message("Done")
  k = k + 1
}

# Merging
paneldata <- merge(paneldata, coef, by = c("timeid", "firmid"))
paneldata <- paneldata %>% arrange(firmid, timeid)

# 2nd stage
lambdas = c()
for(i in min(paneldata$timeid):max(paneldata$timeid)){
  fit = c()
  tmp <- paneldata %>% filter(timeid == i)
  fit <- lm(PF.RF ~ B_Mkt.RF + B_SMB + B_HML + B_RMW + B_CMA, data = tmp)
  lambdas <- rbind(lambdas, cbind(t(fit$coefficients)))
  progress(i, max.value = length(unique(paneldata$timeid)))
  if (i == length(unique(paneldata$timeid))) message("Done")
}

# Statistics
reg <- lm(PF.RF ~ B_Mkt.RF + B_SMB + B_HML + B_RMW + B_CMA,
          data = paneldata)
N <- format(nrow(paneldata), big.mark = ",")
R2 <- round(summary(reg)$r.squared, 3)
R2_adjusted <- round(summary(reg)$adj.r.squared, 3)

tmp3 = c()
for(i in 1:6){
  coefs <- round(mean(lambdas[, i]), 3)
  t.values <- round((mean(lambdas[, i])) / (sd(lambdas[, i]) / sqrt(length(lambdas[, i]))), 3)
  t.values <- paste0("(", t.values,")")
  tmp3 <- rbind(tmp3, coefs, t.values)
}

rownames(tmp3) <- c("Constant", "t", "lambda_RM-RF", "t",
```

```
                    "lambda_SMB","t", "lambda_HML", "t", "lambda_RMW", "t",
                    "lambda_CMA", "t")

# Create a table
stargazer(tmp3,
          header = FALSE,
          rownames = TRUE,
          colnames = TRUE,
          type = "latex")

#=========================================================
#FAMA & MACBETH REGRESSIONS (FAMA & FRENCH 25 PORTFOLIOS)
#=========================================================

# Sentiments (non-lagged)
# Panel format
paneldata <- melt(portfolios_backup, id.vars = "Date")

# Add Fama & French five factors and excess return
paneldata <- merge(paneldata, rs_factors, by = "Date")
paneldata$PF.RF <- paneldata$value - paneldata$RF

# Add sentiment indices
paneldata <- merge(paneldata, S_list, by = "Date" )

# Omit NAs
paneldata <- na.omit(paneldata)
rownames(paneldata) <- NULL

# Add unique firm and month indices
uniqid <- unique(paneldata$variable) %>% sort() %>% data.frame("variable" = .) %>%
  mutate(firmid = seq(1, length(unique(paneldata$variable)), 1))

paneldata <- merge(paneldata, uniqid, by = "variable")

uniqid <- unique(paneldata$Date) %>% sort() %>% data.frame("Date" = .) %>%
  mutate(timeid = seq(1, length(unique(paneldata$Date)), 1))

paneldata <- merge(paneldata, uniqid, by = "Date")

# Arrange data
paneldata <- arrange(paneldata, firmid, timeid)

# 1st stage
minwind = 60

cl <- makeCluster(detectCores(all.tests = TRUE, logical = FALSE))
registerDoParallel(cl)

coef = c()
betanames = paste("B_", c("S", factornames[1:5]), sep = "")
k = 1
for(i in unique(paneldata$firmid)){
  tmp <- filter(paneldata, firmid == i)
  coef_tmp <- foreach(j = (minwind + 1):max(paneldata$timeid), .combine = rbind) %
      dopar% {
    regtemp <- tmp[tmp$timeid >= j - minwind & tmp$timeid <= j - 1, ]
```

```
947     S <- regtemp$S1 # Change (S1, S2, S3, S4)
948     coef_tmp <- lm(PF.RF ~ S + Mkt.RF + SMB + HML + RMW + CMA, data = regtemp)$
            coefficients
949     coef_tmp <- data.frame(t(coef_tmp[-1]), j, i)
950   }
951   coef <- rbind(coef, coef_tmp)
952   if(k == length(unique(paneldata$firmid))){
953     coef = as.data.frame(coef)
954     colnames(coef) = c(betanames, "timeid", "firmid")
955   }
956   coef_tmp = c()
957   progress(k, max.value = length(unique(paneldata$firmid)))
958   if (i == length(unique(paneldata$firmid))) message("Done")
959   k = k + 1
960 }
961
962 # Merging
963 paneldata <- merge(paneldata, coef, by = c("timeid", "firmid"))
964 paneldata <- paneldata %>% arrange(firmid, timeid)
965
966 # 2nd stage
967 lambdas = c()
968 for(i in min(paneldata$timeid):max(paneldata$timeid)){
969   fit = c()
970   tmp <- paneldata %>% filter(timeid == i)
971   fit <- lm(PF.RF ~ B_S + B_Mkt.RF + B_SMB + B_HML + B_RMW + B_CMA, data = tmp)
972   lambdas <- rbind(lambdas, cbind(t(fit$coefficients)))
973   progress(i, max.value = length(unique(paneldata$timeid)))
974   if (i == length(unique(paneldata$timeid))) message("Done")
975 }
976
977 # Statistics
978 reg <- lm(PF.RF ~ B_S + B_Mkt.RF + B_SMB + B_HML + B_RMW + B_CMA,
979           data = paneldata)
980 N <- format(nrow(paneldata), big.mark = ",")
981 R2 <- round(summary(reg)$r.squared, 3)
982 R2_adjusted <- round(summary(reg)$adj.r.squared, 3)
983
984 tmp3 = c()
985 for(i in 1:7){
986   coefs <- round(mean(lambdas[, i]), 3)
987   t.values <- round((mean(lambdas[, i])) / (sd(lambdas[, i]) / sqrt(length(lambdas
          [, i])))), 3)
988   t.values <- paste0("(", t.values,")")
989   tmp3 <- rbind(tmp3, coefs, t.values)
990 }
991
992 rownames(tmp3) <- c("Constant", "t", "lambda_S", "t", "lambda_RM-RF", "t",
993                     "lambda_SMB","t", "lambda_HML", "t", "lambda_RMW", "t",
994                     "lambda_CMA", "t")
995
996 # Create a table
997 stargazer(tmp3,
998           header = FALSE,
999           rownames = TRUE,
1000          colnames = TRUE,
1001          type = "latex")
```

```
1002
1003 # Lagged sentiments
1004 # Panel format
1005 paneldata <- melt(portfolios_backup, id.vars = "Date")
1006
1007 # Add Fama & French five factors and excess return
1008 paneldata <- merge(paneldata, rs_factors, by = "Date")
1009 paneldata$PF.RF <- paneldata$value - paneldata$RF
1010
1011 # Add lagged sentiment indices
1012 paneldata <- merge(paneldata, S_list_lag1, by = "Date")
1013
1014 # Omit NAs
1015 paneldata <- na.omit(paneldata)
1016 rownames(paneldata) <- NULL
1017
1018 # Add unique firm and month indices
1019 uniqid <- unique(paneldata$variable) %>% sort() %>% data.frame("variable" = .) %>%
1020   mutate(firmid = seq(1, length(unique(paneldata$variable)), 1))
1021
1022 paneldata <- merge(paneldata, uniqid, by = "variable")
1023
1024 uniqid <- unique(paneldata$Date) %>% sort() %>% data.frame("Date" = .) %>%
1025   mutate(timeid = seq(1, length(unique(paneldata$Date)), 1))
1026
1027 paneldata <- merge(paneldata, uniqid, by = "Date")
1028
1029 # Arrange data
1030 paneldata <- arrange(paneldata, firmid, timeid)
1031
1032 # 1st stage
1033 minwind = 60
1034
1035 cl <- makeCluster(detectCores(all.tests = TRUE, logical = FALSE))
1036 registerDoParallel(cl)
1037
1038 coef = c()
1039 betanames = paste("B_", c("S", factornames[1:5]), sep = "")
1040 k = 1
1041 for(i in unique(paneldata$firmid)){
1042   tmp <- filter(paneldata, firmid == i)
1043   coef_tmp <- foreach(j = (minwind + 1):max(paneldata$timeid), .combine = rbind) %
         dopar% {
1044     regtemp <- tmp[tmp$timeid >= j - minwind & tmp$timeid <= j - 1, ]
1045     S <- regtemp$S1_lag1 # change (S1_lag1, S2_lag1, S3_lag1, S4_lag1)
1046     coef_tmp <- lm(PF.RF ~ S + Mkt.RF + SMB + HML + RMW + CMA, data = regtemp)$
           coefficients
1047     coef_tmp <- data.frame(t(coef_tmp[-1]), j, i)
1048   }
1049   coef <- rbind(coef, coef_tmp)
1050   if(k == length(unique(paneldata$firmid))){
1051     coef = as.data.frame(coef)
1052     colnames(coef) = c(betanames, "timeid", "firmid")
1053   }
1054   coef_tmp = c()
1055   progress(k, max.value = length(unique(paneldata$firmid)))
1056   if (i == length(unique(paneldata$firmid))) message("Done")
```

```
  k = k + 1
}

# Merging
paneldata <- merge(paneldata, coef, by = c("timeid", "firmid"))
paneldata <- paneldata %>% arrange(firmid, timeid)

# 2nd stage
lambdas = c()
for(i in min(paneldata$timeid):max(paneldata$timeid)){
  fit = c()
  tmp <- paneldata %>% filter(timeid == i)
  fit <- lm(PF.RF ~ B_S + B_Mkt.RF + B_SMB + B_HML + B_RMW + B_CMA, data = tmp)
  lambdas <- rbind(lambdas, cbind(t(fit$coefficients)))
  progress(i, max.value = length(unique(paneldata$timeid)))
  if (i == length(unique(paneldata$timeid))) message("Done")
}

# Statistics
reg <- lm(PF.RF ~ B_S + B_Mkt.RF + B_SMB + B_HML + B_RMW + B_CMA,
          data = paneldata)
N <- format(nrow(paneldata), big.mark = ",")
R2 <- round(summary(reg)$r.squared, 3)
R2_adjusted <- round(summary(reg)$adj.r.squared, 3)

tmp3 = c()
for(i in 1:7){
  coefs <- round(mean(lambdas[, i]), 3)
  t.values <- round((mean(lambdas[, i])) / (sd(lambdas[, i]) / sqrt(length(lambdas
      [, i])))), 3)
  t.values <- paste0("(", t.values,")")
  tmp3 <- rbind(tmp3, coefs, t.values)
}

rownames(tmp3) <- c("Constant", "t", "lambda_S", "t", "lambda_RM-RF", "t",
                    "lambda_SMB","t", "lambda_HML", "t", "lambda_RMW", "t",
                    "lambda_CMA", "t")

# Create a table
stargazer(tmp3,
          header = FALSE,
          rownames = TRUE,
          colnames = TRUE,
          type = "latex")
```