Stockholm School of Economics
Department of Finance

Master's Thesis – Fall 2015
Tutor: Anders Anderson

Investor Sentiment and Stock Returns

ERIK LUNDELL[α] & SHUANGYU PENG[β]

ABSTRACT

We examine whether investor sentiment can predict next-day stock returns and if the predictive characteristic differs when discriminating between positive and negative sentiment. We propose two novel sentiment proxies derived from a unique dataset using machine learning algorithms to approximate sentiment. In a sample of 286 Swedish equities from February 2014 to June 2015, we find that (1) investor sentiment can predict next-day stock returns; (2) the negative sentiment has a larger absolute economic effect than positive sentiment; and (3) an indication of that a simple sentiment-based trading strategy can earn substantial risk-adjusted returns.

Keywords

Investor sentiment, return predictability

Acknowledgements

We thank our thesis tutor Anders Anderson for his guidance and genuine interest in the study. We thank Tomas Tysén, Hans Möller and the team at Modular Streams for providing us with the unique sentiment dataset.

[α] 22084@student.hhs.se
[β] 40805@student.hhs.se

1 INTRODUCTION

The efficient market hypothesis assumes that stock prices reflect all publically available information and that returns are derived from carrying risk.[1] Put differently, the market value of a firm should equal to the expected discounted value of future cash flows of that firm conditional on investors' information set.[2] Therefore, investor sentiment should not affect stock market returns. However, the perception is quite the opposite in behavioural finance. There are models showing for example that noise traders can cause stock prices to deviate from their fundamental value.[3] Also, empirical findings from psychology studies, studying how market participants form beliefs about stocks, support the existence of a relationship between investor sentiment and stock returns.[4]

The recognition of the relationship between investor sentiment and stock returns extends into the professional finance community as well. For example, Value Line[5] provides relative stock rankings and research reports in which momentum and investor sentiment are taken into account among other factors. Furthermore, in September 2015, Bloomberg, the primary financial data provider in the industry, signed an agreement with Twitter regarding the use of finance-related content on Twitter in the Bloomberg terminal. This can be linked to Bollen et al. (2011), which document significant correlation between the Dow Jones Industrial Average (DJIA) and a proxy for public sentiment, entirely based on text analysis of Twitter posts ("tweets"). Even though investor sentiment has gained significant traction over the past couple of years and earned a level of legitimacy both in the academic and professional community, the topic remains much debated as academic studies continue to show contradicting results.

We examine the effect of investor sentiment on stock returns by constructing two sentiment proxies, incorporating computer sciences methods in the collection and approximation of investor sentiment, and intend to answer the following research questions:

- *Does investor sentiment help to predict next-day stock returns?*
- *Do the predictive characteristics differ between positive and negative investor sentiment?*
- *Does a sentiment-based trading strategy deliver risk-adjusted returns?*

---

[1] Fama (1970)
[2] Malkiel (1992)
[3] Shleifer and Summers (1990)
[4] Barberis et al. (1998)
[5] www.valueline.com

We focus our study on the Swedish stock market, covering a total of 286 companies, in the time period from February 2014 to June 2015. We base our statistical analysis on a unique dataset obtained in part from the information technology company Modular Streams. The company collects the raw sentiment dataset from a range of online sources and analyses it through the use of machine learning algorithms. The remaining data is collected from a range of acknowledged financial data sources.

A common challenge and source of discussion in the literature is the construction of an accurate and robust proxy for investor sentiment, as investor sentiment cannot be directly measured. Throughout previous research, a range of different proxies is presented. To name a few, Neiderhoffer (1971) constructs a discrete proxy based on the font size of New York Times headlines; Antweiler and Frank (2004) measure the level of disagreement of message board posts and post volume when constructing a sentiment proxy; Brown and Cliff (2004, 2005) use survey data collected from market participants. We propose two novel sentiment proxies based on our unique sentiment dataset, namely: Sentiment Score (*SSc*) and Sentiment Shock (*SSh*). *SSc* is a discrete proxy that aims to capture both positive and negative investor sentiment. *SSh* consists of a pair of dummy variables (*Pos_SSh* and *Neg_SSh*) that discriminates between positive and negative sentiment and aims to capture events when the underlying sentiment experiences a shock.

We use OLS regressions for our main empirical tests and document statistically significant evidence that investor sentiment predicts next-day stock returns given a positive relationship between *SSc* and next-day excess returns and abnormal returns. As we expected, the positive relationship between investor sentiment and next-day stock returns is reaffirmed when discriminating between positive and negative sentiment, using the proxies *Pos_SSh* and *Neg_SSh*. The coefficients in our regression results are positive (negative) for the corresponding positive (negative) sentiment dummy variable. Further, negative sentiment shocks predict on average a larger absolute effect on next-day stock returns than positive sentiment shocks. Our results confirm existing findings in, among others, Tetlock (2007) and Tetlock et al. (2008).

To address our last research question, we construct an equally weighted long-short trading strategy based on a daily ranking derived from *SSc*. In the sample period, our portfolio yields a risk-adjusted return similar to the magnitude documented by Tetlock et al. (2008), under the assumption of perfect financial markets.[6] However, regression results are not statistically significant. Additionally, the profits are quickly eroded when taking reasonable transaction costs into account.

[6] No transaction costs, capital gains tax, etc.

Our findings complement existing research in a number of ways. The documented findings largely confirm results in previous studies within behavioural finance, contradicting the efficient market hypothesis. Although our proxies are different from proxies used in previous studies, we document similar findings. The results suggest that *SSc*, *Pos_SSh* and *Neg_SSh* are similar to other proxies when approximating investor sentiment. Furthermore, this endorses the legitimacy of the application of data mining and machine learning methods for data collection and sentiment analysis. Additionally, the difference in characteristics existing in positive and negative sentiment shocks highlights the need for discriminating proxies, in terms of separating positive and negative sentiment, instead of one-sided proxies.

Regarding the future applicability of our results and sentiment proxies, it is a possibility to extend classic asset pricing models, as suggested by Baker and Wurgler (2007), to include factors of investor sentiment instead of only accounting for commonly used risk factors.[7] Also, our sentiment proxies can be useful in testing the predictions of theoretical models, *e.g.* Barberis et al. (1998), which predicts that one-time strong news events should generate an overreaction, however does not show any real evidence supporting their model's prediction.

The remainder of the paper is structured as follows: **Section 2** reviews existing literature in the field of investor sentiment and the technical background of the paper. **Section 3** describes our dataset and the collection and proxy construction methods we use. **Section 4** describes the empirical methods we use to address our research questions. **Section 5** reports and discusses our findings. **Section 6** presents concluding remarks and discusses possible future research ideas.

[7] Carhart (1997). Including market risk, size (*SMB*), book-to-market (*HML*), momentum (*UMD*)

2 LITERATURE REVIEW

Before delving into our research, we look at existing literature in the field of investor sentiment. The section starts with a review of the most directly relevant papers and moves on to exploring literature covering related anomalies to the efficient market hypothesis. Understanding common methods, previous findings and the general context of the research topic is beneficial in order to understand our methods and results. We end the section with a short review of the technical background for the understanding of our sentiment dataset, as unconventional data collection methods are used.

2.1 Investor sentiment

Research on investor sentiment lies under the field of behavioral finance and is somewhat different to classic research relating to the efficient market hypothesis ("EMH"). However, it overlaps with mainstream research as the analysis and arguments often are influenced by research relating to anomalies challenging the EMH. Following is a short historic review of this field of research which culminates in the review of our main reference papers, Tetlock (2007) and Tetlock et al. (2008).

One of the first studies exploring this subject is Niederhoffer (1971) as he examines world events' effect on stock prices through the analysis of newspaper articles. His findings show a strong tendency for price changes on the first and second day following world events (in the same direction as the sentiment). He also documents a reversal following extremely bad news in the subsequent 2-5 days suggesting the market overreacts to bad news. An important distinction to make is whether world events drive stock returns or whether news articles drive stock returns. In Niederhoffer (1971) this is not addressed as news articles are used as a proxy for world events. However, this distinction is crucial to make in the body of research covering news and sentiment that really started to take off in the 90's. Almost two decades after Niederhoffer (1971), Roll (1988) among others provides evidence that stories in the financial press do not affect stock prices, suggesting news articles contain little new informational value and do not drive stock returns.[8] Contradicting Roll's findings, Mitchell and Mulherin (1994) document evidence (although weak) that Dow Jones announcements and aggregated measures of market activity are directly related.[9] Chan (2003) examines monthly stock returns and documents a drift (up to twelve months) following public news events. The effect is strongest for smaller firms in combination with bad news events however not exclusive to these conditions. Pritamani and Singal (2001) find similar results for a small

---

[8] See also Cutler et al. (1989). Findings show that news stories unaccompanied by quantitative macroeconomic events do not help to explain stock price movements.

[9] Aggregated measures of market activity include price volatility, trading volume and stock returns.

subset of stocks from NYSE/AMEX in a sample from 1990 to 1992. They document both positive and negative stock price drifts for up to 20 days following news events in the Wall Street Journal and Dow Jones News Wire.

Womack (1996) takes a different approach and examines if analyst recommendations have investment value. He finds supporting evidence that post recommendation stock returns are significantly positive (negative) after buy (sell) recommendations. This is particularly interesting to this thesis when considering the information signal involved. An analyst recommendation announcement is a change of *opinion* by a market participant, while other news announcement may be new public facts, *e.g.* quarterly earnings announcements. Bagnoli et al. (1999) compare analyst's earnings forecasts and so called whisper forecasts and find that whisper forecasts tend to be more accurate despite the "unofficial nature" of the forecasts.[10] Further, Wysocki (1999) examines message-posting volume on stock message boards on the web and finds that next-day returns can be predicted in his sample of 50 stocks with the highest posting volume between January and August 1998. The latter study is one of the first that examines the predictability of online-based message activity on measures of market activity.

In the 00's many similar studies emerged covering online chat forums, media sentiment, social media, and search engine analysis.[11] Antweiler and Frank (2004) thoroughly investigate the relationship between stock message board activity (on Yahoo! Finance and Raging Bull) and stock returns, trading volume, and price volatility.[12] They find significant evidence that; high message board volumes predict negative next-day stock returns, although economically small; disagreements among posted messages predict an increase in subsequent trading volume, however they document a reversal on the next trading day; lastly, message posting helps to predict price volatility.

Brown and Cliff (2004, 2005), Baker and Wurgler (2006, 2007), and Livnat and Petrovits (2009) study the relationship between investor sentiment and stock returns. The studies differ in the proxies of sentiment as Brown and Cliff (2004, 2005) use survey data while Baker and Wurgler (2006, 2007) and Livnat and Petrovits (2009) use a constructed sentiment index.[13] All but Brown and Cliff (2004) find supportive evidence that investor sentiment helps to predict future stock returns. They find a negative relationship between investor sentiment and subsequent stock returns, suggesting that assets prices deviate from

---

[10] Whisper forecasts are unofficial earnings forecasts circulating among traders and investors.

[11] Joseph et al. (2011) document evidence that, in a sample from 2005-2008 of S&P500 stocks over weekly horizon, online search intensity reliably predicts abnormal stock returns and trading volume. Da et al. (2011) find agreeing evidence from a sample from 2004-2008 of Russell 3000 stocks, further specifying a 2 week abnormal return continuation and a subsequent price reversal within a year.

[12] The study also provides a good mapping of early 2000's research in this field.

[13] Developed by Baker and Wurgler (2006), derived through a principal component analysis of the six most-commonly used proxies for sentiment: closed end fund discount, NYSE share turnover, number of IPOs, first-day returns on IPOs, equity share of new equity and debt issues, and dividend premium.

their intrinsic value in times of high/low sentiment and subsequently revert back to fundamentals. Livnat and Petrovits (2009), along with other current studies, take a slightly different approach but reaches similar conclusions.[14] Livnat and Petrovits look at immediate and long-run market reactions to earnings announcements following periods of high/low sentiment. They find that the post-earnings announcement drift ("PEAD") phenomenon is amplified when sorting on sentiment prior to the announcement, *i.e.* upward (downward) PEAD following positive (negative) earnings surprises is greater in periods of low (high) investor sentiment.

An author that has taken influence from the above regarding media coverage, message board activity, investor sentiment and the stock market, referencing for example Cutler, Poterba, and Summers (1989), Antweiler and Frank (2004), and Baker and Wurgler (2006) among others, is Paul Tetlock. Tetlock (2007) studies the interaction between the media and the stock market (Dow Jones index) by constructing a measure of media pessimism using daily content from the Wall Street Journal. His findings robustly rejects the hypothesis that media is an irrelevant noisy variable with no relation to stock markets. He further documents that high levels of media pessimism robustly predicts downward pressure on stock prices, followed by a reversion to fundamentals in the next few days. This suggests that media content neither contains any new information nor can be used as a proxy for new information about fundamentals. However in Tetlock et al. (2008) the latter conclusion is somewhat contradicted. The paper documents evidence that news stories prior to earnings announcements contain useful information about otherwise hard-to-quantify fundamentals, consequently reliably predicting both future earnings and stock returns. Further, the study extends Tetlock (2007) by examining individual firms' stock returns opposed to index returns. Again, statistical evidence supports the hypothesis that media content is a helpful predictor of future stock returns. By testing the interaction between negative words in firm-specific news stories and next-day stock returns, Tetlock et al. (2008) find a slight underreaction as abnormal returns in day $t+1$ is significantly negative (although minor compared to initial reaction on day t).

[14] Bird and Yeung (2012) and Bird et al. (2014) examine the combined impact of market uncertainty and market sentiment on how investors react to information.

Table I
Review of sentiment proxies

This table shows a summary review of the sentiment proxies, sample periods and sample geographies used in previous literature.

Paper	Sample period	Sample geography	Sentiment Proxies
Niederhoffer (1971)	1960-1965	U.S.	Now York Times headlines, including size and 7-point good-bad scale; market-wide
Cutler et al. (1998)	1926-1985	U.S.	Macroeconomic news; market-wide
Chan (2003)	1980-1999	U.S.	Public news; stock-level
Antweiler and Frank (2004)	2000	U.S.	Messages in Yahoo! Finance and Raging Bull; stock-level
Brown and Cliff (2004, 2005)	1965-1998; 1963-2000	U.S.	Surveys and newsletters; market-wide
Lemmon and Portniaguina (2006)	1986-2006	U.S.	Consumer confidence index; market-wide
Kummer and Lee (2006)	1991-1998	U.S.	Retail investors transactions; stock-level
Baker and Wurgler (2006, 2007)	1962-2001	U.S.	Sentiment index constructed based on principal component analysis of close-end fund discount, turnover, numbers and first-day returns of IPOs, etc.; market-wide
Livnat and Petrovits (2007)	1987-2005	U.S.	Similar to Baker and Wurgler (2007)
Tetlock (2007)	1984-1999	U.S.	Content of Wall Street Journal; stock-level
Tetlock et al. (2008)	1980-2004	U.S.	Fraction of negative words in news articles; stock-level
Kaniel et al. (2008)	2000-2013	U.S.	Buing and selling volumes by individual investors; stock-level
Baker et al. (2012)	1980-2005	Canada, France, Germany, Japan, U.K., U.S.	Sentiment index constructed based on turnover, numbers and first-day returns of IPOs; market-wide

2.2 Efficient market hypothesis and related anomalies

Like the majority of financial studies performed during the past half century, we refer to the efficient market hypothesis (EMH) and the associated empirical work written by Fama in 1970. It has provided organizing principles for empirical financial economics since its introduction, and continues to do so yet today. It extends the research of the random walk hypothesis[15] adding new dimensions to the definition of market efficiency, namely semi-strong[16] and strong efficiency. However, like its predecessor, the main point remains, stock price movements are unpredictable.

Following Fama's contributions, a vast plethora of literature emerged challenging the EMH. In fact, even Eugene Fama states that markets are not efficient. Efficiency is an ideal

[15] First seen in a primitive state in the paper Théorie de la Spéculation (The Theory of Speculation) by Louis Bachelier, published in 1900. Since then the theory has been further developed by many, culminating in a frequently cited paper by Fama (1965).

[16] First stage testing of semi-strong efficiency had been performed earlier by Fama et al. (1969) regarding announced stock splits, Ball and Brown (1968) regarding earnings announcements, and Scholes (1969) regarding new stock issuance and secondary offerings. All of which supported semi-strong market efficiency.

that real-world markets only can approach. Empirical work can only find how close to or far from the ideal a given market is. Thus it is evident that there will always be anomalies under specific circumstances and conditions. Some anomalies are more renown than others, having gained a lot of traction in both the academic community as well as the professional.

2.2.1 Momentum

Momentum is one of the largest research themes regarding anomalies in the EMH. One of the first studies to challenge the early stage EMH or more correctly the random walk hypothesis was Levy (1967). He states, "it appears that superior profits can be achieved by investing in securities which historically have been relatively strong in price movement", calling this strategy *relative strength*, *i.e.* price momentum. Naturally, his conclusions were immediately met both by scepticism[17] and other studies supporting his findings.[18]

Recurring themes in the efforts of deriving the momentum anomaly as well as being prominent research fields on their own are market under- and overreaction. These are also directly related to research on investor sentiment and behavioural finance.

2.2.2 Overreaction

De Bondt and Thaler (1985, 1987) document, in contrast to Jegadeesh and Titman (1993, 2001) and in violation on Bayes' rule, return reversals over long-term horizons, advocating that a contrarian trading strategy will yield abnormal returns. Return reversals are also documented over short-term horizons by Lehmann (1990) and Jegadeesh (1990), forming portfolios with week-long and month-long holding periods respectively. Like others, De Bondt and Thaler explain return reversals as a correction of investor overreaction to earnings information.[19]

The field of overreaction continues to interest researchers throughout the 90's and 00's. However, no clear consensus in academia is reached on the reasons for or even the existence of overreaction.[20]

---

[17] Jensen (1967); Lo and MacKinlay (1990) suggesting a delay to common factors.
[18] Grinblatt and Titman (1989, 1993) examining mutual fund performance; Jegadeesh and Titman (1993), a benchmark paper within this field, provides a relative strength, *price momentum*, trading strategy, later supported by Jegadeesh and Titman (2001) and Chan et al. (1999); Rouwenhorst (1998) finds that momentum is robust to internationally diversified portfolios; Moskowitz and Grinblatt (1999) controlling for *industry momentum*; Cohen and Frazzini (2008) and Menzly and Ozbas (2006) controlling for *costumer momentum*.
[19] See for example Chopra et al. (1992).
[20] Chopra et al. (1992) suggest return reversals are due to correction of investor's overreaction to earnings information; Lo and MacKinlay (1990) argue for a delayed price reaction to common factors rather than overreaction; Jegadeesh and Titman (1993) attribute the results to the *January effect*; Abarbanell and Bernard (1992) reject any link between return reversals and overreaction; Fama and French (1992, 1996) argue that contrarian strategies are fundamentally risker, which is later contradicted by Lakonishok et al. (1994) suggesting such strategies exploit suboptimal behavior of investors.

2.2.3 Underreaction

Closely related to overreaction is the field of underreaction. It is also commonly discussed in the presence of the post-earnings announcement drift ("PEAD"). Ball and Brown (1968) were the first to show evidence of a drift in stock prices after an earnings announcement. Cumulative abnormal returns drift upward for good news firms and downward for bad news firms. Later studies have found confirming evidence. [21]

Competing explanations for PEAD has previously fallen into one of two categories namely: at least a portion of the price response to new information is delayed due to transaction costs or investor's inability to assimilate available information; or the CAPM (used to calculate abnormal returns) is incomplete or misestimated, *i.e.* the abnormal returns documented is actually the fair compensation for bearing risk. However, a frequently cited paper, Bernard and Thomas (1989), examines the two and find no evidence that can plausibly reconcile with the latter, however is consistent with the argument for a delayed price response.

In the late 90's and 00's more studies examine investor phycology and joint models for explaining under- and overreaction. Barberis et al. (1998), Daniel et al. (1998), and Hong and Stein (2007), present behavioral models that are based on the idea that momentum profits arise because of inherent biases in the way that investors interpret information.

Daniel et al. (1998) conclude that, "in contrast with the common correspondence of positive (negative) return autocorrelations with underreaction (overreaction) to new information, we show that positive return autocorrelations can be a result of continuing overreaction. This is followed by long-run correction. Thus, short-run positive autocorrelations can be consistent with long-run negative autocorrelations." Bird and Yeung (2012) document that investors overreact to bad news and underreact to good news to a degree positively correlated with uncertainty in the market. [22]

2.3 Technical background

In our research, we propose two original proxies of investor sentiment – Sentiment Score (*SSc*) and Sentiment Shock (*SSh*). They are collected and computed with the help of data mining, machine learning and sentiment analysis. We briefly introduce the definitions and functions of these methods. We further explain the rationale behind our data collection method in *Section 3.2*.

Data mining (or "*web-scraping*") is the first step in collecting the sentiment data used in our research. Data mining is the process of extracting large quantities of data from

[21] Foster et al. (1984) document a 60-day drift in stock prices subsequent to an earnings announcement. Also the magnitude of the drift is positively correlated to the magnitude of the unexpected earnings change and negatively correlated to firm size, but not exclusive to small firms.
[22] See also Francis et al. (2007).

online sources in order to construct a dataset. The extracted data is based on a range of specified criteria and sources. The collection criteria can include specific key words, topics, date of the content and so on and so forth. The sources can range from online encyclopaedias to chat forums or Twitter posts and even video content. As we are interested in investor sentiment relating to individual stocks, we extract sections of text relating to listed stocks based on specific search criteria but from a broad range of sources including newspapers, stock forums and message boards etc. (a selection of sources and criteria can be found in *Section 3.2*).

The collected data is then processed using a machine learning method derived from algorithmic programming. Samuel (1959) defines machine learning as a "field of study that gives computers the ability to learn without being explicitly programmed". As mentioned, we are interested in the sentiment (or the optimism/pessimism) of the text in every data point collected through the data mining. Because of the large quantity of data collected, it is unrealistic to analyse and judge the sentiment manually. A machine learning algorithm teaches computers to automatically perform the analysis and judge the underlying sentiment in every data point in our dataset. Machine learning is a method that of course can be used for many different purposes. However, our application is to perform the sentiment analysis and make a judgement whether the sentiment of the texts is positive, negative or neutral.

3 DATA

This section describes our dataset, sources and methods used for the collection and initial processing of the dataset. The data is categorized as financial data and sentiment data. The former includes stock price, stock turnover, market return, risk-free rate and risk factor data with a daily frequency and ranges from January 2010 to June 2015. The latter includes our unique dataset with our proxies for investor sentiment with a daily frequency and ranges from February 2014 to June 2015. The final merged dataset used for the statistical analysis is presented in *Section 3.4*.

3.1 Financial data

We obtain stock prices, daily turnover and market index data from Nasdaq Stockholm. The OMX Stockholm All-Share Cap GI (*OMXSCAPGI*) index is used as the proxy for market returns as it contains 300 equities ranging from small to large cap equities, thus the most representative for the overall market performance and most comparable to our list of sample stocks (refer to *Section 3.3*). The risk-free rate is derived from the over-night Stockholm Interbank Offered Rate (*STIBOR*), collected from Sveriges Riksbank, as it best matches the maturity/holding period of our trading strategy. Two of the Fama-French (1993) factors (*SMB* and *HML*) and the momentum factor (*UMD,* Carhart (1997)) are collected from the Kenneth R. French Data Library.[23] The original factor data is computed based on equities in the United States as daily Swedish factors are not readily available through any distinguished source. Thus, we convert the US factors to approximated Swedish factors, by using the method advocated by Calvet et al. (2007), in order to match our otherwise exclusively Swedish dataset. The method effectively adjusts the factor return to the FX risk of investing in foreign portfolios (denominated in USD) with domestic currency (SEK) as the base currency. The SEK/USD exchange rates are collected from FactSet.

When merging the financial data some observations are excluded due to a mismatch between the US factor data and Swedish market data (stock prices, turnover, and market return). This is due to for example, a mismatch between US stock market holidays and Swedish stock market holidays or missing data for specific stocks due to ceased trading (regulated by Nasdaq). The summary statistics of the financial dataset are presented in Table II below.

[23] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Table II
Summary statistics – financial data

This table shows the summary statistics for financial data corresponding to our 286 sample stocks for the time period January 2010 to June 2015.

Variable	Obs	Mean	Std. Dev.	Min	Max
Return	364,783	0.0007	0.0388	-0.9031	8.7794
Turnover (in MSEK)	364,783	35.20	122.00	0.00	8,070.00
Market Return	364,783	0.0006	0.0115	-0.0619	0.0631
SMB	355,245	0.0001	0.5260	-0.0200	0.0369
HML	355,245	-0.0001	0.4109	-0.0166	0.0170
UMD	355,245	0.0002	0.6110	-0.0264	0.0252
Risk-free Rate	364,783	0.0000	0.0000	0.0000	0.0001

3.2 Sentiment data

The sentiment data is central for this thesis and is a source of debate throughout existing research. Because investor sentiment cannot be directly observed, different approximation methods have been used in the past in order to attain a good proxy for investor sentiment. We incorporate computer science (data mining and machine learning) when constructing our proxies for investor sentiment (Sentiment Score (*SSc*) and Sentiment Shock (*SSh*)), further described in the following sub-sections. We construct two proxies in order to better investigate our research questions and ensure that our results remain robust across different measures of sentiment.

The original dataset obtained from Modular Streams consists of 92,313 data points, collected from news websites, Twitter, YouTube, Facebook, Google+ and RSS feeds (a selection of sources are presented in Appendix I). In total, the dataset includes data from 492 companies on Nasdaq Stockholm, Nasdaq First North Stockholm and AktieTorget. However, as described in *Section 3.3*, we exclude a portion of our original data as result of quality control measures.

3.2.1 Data collection and approximation of sentiment

The collection of the sentiment data and the sentiment approximation is conducted by the information technology company Modular Streams. They use a combination of data mining and machine learning methods in order to collect and assign a sentiment to the data points. We refer to each data point as a sentiment observation, which relates to for example, a news article or a tweet about a specific stock. Note that there can be several sentiment observations per stock per day. We later construct our sentiment proxies based on the sentiment dataset, further described in the following sub-sections (*Section 3.2.2* and *Section 3.2.3*). Here follows a short description of the process used by Modular Streams:

The first step is to extract data from a range of online sources using data mining. Sections of text are extracted (maximum 160 words per section) based on specific extraction criteria, which works as a filter. The filter is based on three selection criteria: the first searches for a company name, ticker or common company name variations (*e.g.* Alfa Laval, ALFA, Alfa Laval's) in order to sort the data after specific companies; the second searches for key words relating to stock information, earnings or other words relating to company financials (*e.g.* when searching for the company Boliden or Sandvid, which are also Swedish towns, the filter excludes search results not relating to the actual companies); the third applies a machine learning algorithm that determines the relevance of text (*e.g.* when extracting sections of a news article, only relevant sections are extracted so the subsequent sentiment analysis is not distorted).[24] Stocks with generic or dual-meaning company names or tickers are excluded in the first filter, as this can create unwanted noise in the search results (*e.g.* the company Avanza Bank Holding is excluded as "Avanza" is often used in the context of online trading, mostly not referring to the company or the company's performance).[25]

The next step is to perform sentiment analysis on texts and judge whether the sentiment of the underlying text is positive, negative or neutral. This step is done by the use of a so-called machine learning algorithm. A note worth mentioning is that within sentiment analysis, a binary position is usually assumed: either for or against, positive or negative and so on and so forth. However, thanks to Modular Streams' advanced algorithm, a three-point scale grading can be assigned: positive, negative or neutral. In our opinion, this indicates a more realistic representation of investor sentiment.

Before the algorithm works automatically, it has to be taught how to judge the sentiment in the data points. Modular Streams hires a diverse set of individuals to manually judge, according to their opinion, whether a text is perceived as positive, negative or neutral. It is important to use a number of individuals at this stage as personal attitudes are captured and reflected in the algorithm, thus the average of opinions is more representative for the opinion of the entire population.[26] This is done for 3000-5000 texts in order ensure the quality of the algorithm's output. At this stage, additional filters and techniques are necessary in order to reduce the noise and increase the quality of the output. However, further specification of these filters is confidential company information and may not be disclosed in this thesis. Modular Streams regularly conducts tests to ensure that the output upholds a minimum level of accuracy when determining sentiment. For the dataset used in this thesis, the accuracy lies at just below 90 percent, which we find satisfactory.

The final step is to compile the output into a useable dataset. The original dataset includes a range of additional information collected by the algorithm (though not used in this thesis), including if the text is general information, an opinion, a question or an answer; if the

---

[24] The first two requirements are similar to the ones in the application of Harvard-IV-4 psychosocial dictionary in WSJ and DJNS in Tetlock et al. (2008).
[25] Da et al. (2011)
[26] This is similar to Niederhoffer (1971), which hires independent observers/volunteers to judge the information of New York Times headlines on a good-bad scale.

author is professional or private; and of course the sources of the sentiment observations (*e.g.* Twitter, stock message boards, newspapers).

3.2.2 Sentiment score

We define investor sentiment as the general optimism and pessimism relating to a specific stock. This cannot be directly observed and has to be approximated by a constructed sentiment proxy. We construct two sentiment proxies; the first is called Sentiment Score (*SSc*) and described in this sub-section; the second Sentiment Shock (*SSh*) and is described in the next sub-section. The Sentiment Score can be interpreted as the larger the positive (negative) value of *SSc* is, the larger the optimism (pessimism) is regarding that stock for that day.

We start by assigning each sentiment observation +1, -1 or 0 points for positive, negative or neutral sentiment, respectively. We then sum all sentiment observations by company and date. As such, we construct a discrete variable, with one observation for each stock and for each day in our sample. Note that *SSc* does not reflect the number of sentiment observations per day per company as the sum can be 0 even though multiple observations have been recorded.

The intention of constructing *SSc* is to mimic the proxy used in Niederhoffer (1971). He constructs a sentiment proxy based on New York Times headlines, which are judged on a scale from 1 to 7 (1 reflecting negative sentiment and 7 reflecting positive sentiment) by a panel of individuals. However, while Niederhoffer limits his proxy to integers from 1 to 7, our proxy reflects the sum (per company and day) of the points assigned to each sentiment observation. We argue that this allows our proxy to be more accurate and nuanced compared to a narrowly scaled proxy. Additionally, compared to Neiderhoffer (1971), *SSc* is constructed on stock-level while Neiderhoffer constructs the sentiment proxy reflecting market-wide sentiment, since his research focuses on market (index) returns.

An important decision when sorting the sentiment observations per day (and company) is the time for the cut-off point each day. We set the cut-off point for each day at 17:15 (15 minutes before market closing), thus the observations included in day *t* are observed with a time stamp ranging from 17:15:00 on day *t-1* to 17:14:59 on day *t*. This decision is influenced by Tetlock et al. (2008) which exclude news coverage released 30 minutes before market closing. We argue that 30 minutes is too long as markets today are more efficient and it has become less costly to trade (compared to 7 years ago).[27] Note that for Mondays, all sentiment observations are aggregated from 17:15:00 on Friday to 17:14:59 on Monday. This might distort *SSc* on Mondays as observations are aggregated during several

[27] Different cut-off points are tested in unreported tests, ranging from 17:00 to 17:30 (at closing), and results are fundamentally the same, where neither the sign nor significance of coefficients are affected to any noteworthy degree.

days instead of only one. However by controlling for both weekdays and number of sentiment observations in our regressions, this potential effect is extracted from *SSc* in our results. Further, to mitigate issues with extreme values creating skewed results, we exclude observations when *SSc* exceeds +/-50.[28]

3.2.3 Sentiment shock

Our second sentiment proxy is Sentiment Shock (*SSh*), which is a pair of dummy variables divided into either positive (*Pos_SSh*) or negative (*Neg_SSh*) sentiment shock. As *SSh* separates positive and negative sentiment into two variables, opposed to the proxy *SSc*, results are expected to show if positive and negative sentiment affect stock returns in a similar way, to the same extent or at all. This is derived from our second research question – *Do the predictive characteristics differ between positive and negative investor sentiment?* Also, by constructing the proxy as a pair of dummy variables, different from the discrete variable *SSc*, it serves as a test for the robustness of our results.

The dummy for positive (negative) sentiment shock is assigned 1 if the value of *SSc* exceeds +3 (-3) and assigned 0 if *SSc* does not. We choose the cut-off threshold to be +/-3 when constructing *Pos_SSh* and *Neg_SSh*, as the mean and standard deviation of *SSc* are 0.23 and 1.55 respectively; and +/-3 is approximately two unites of standard deviation from the mean.[29] However, the threshold values are not perfect since *SSc* fails to comply with normal distribution. However, we still expect results from the dummy variables to benefit our analysis and our understanding of the differences in characteristics of optimism and pessimism.

3.3 Sample stocks

To ensure that the quality of our dataset upholds a high standard, we apply certain minimum requirements that rid the dataset from companies that are not desirable from a statistical point of view. The exclusion process is mainly driven by the quality of the financial data even though our first selection is fully determined by the list of companies included in our raw sentiment dataset.

The raw sentiment dataset attained from Modular Streams contains 492 companies listed on Nasdaq Stockholm, Nasdaq First North or AktieTorget. We exclude stocks based on the following exclusion criteria: 1) the stock must trade on Nasdaq Stockholm or Nasdaq First North to ensure a certain minimum size (market capitalization), level of liquidity and reasonable volatility, *i.e.* stocks trading on AktieTorget, an unregulated market, are smaller

[28] Excluded observations with a *SSc* value exceeding +/-50 amount to 7 observations.

[29] Different threshold values are tested in unreported tests, ranging from 3 to 8. Results show a decreasing statistical significance of the related coefficients as the threshold value increases and the number of "triggering" observations decrease.

and generally of lower quality (low liquidity, high volatility, less corporate governance); 2) the stock has to have at least 50 days of trading history.[30] The main reason for this is to ensure that there is sufficient historical data available to accurately estimate the abnormal return for the stock.

By applying our exclusion criteria to our dataset, 206 stocks are excluded and thus 286 stocks remain with high financial and sentiment data quality. A table of all sample stocks is presented in Appendix II.

3.4 Dataset

After collecting all financial and sentiment data, constructing our sentiment proxies and selecting our sample stocks, we merge financial data and sentiment data into one dataset to facilitate further analysis. The final dataset that we conduct all statistical testing and analysis on is presented in the Table III and Table IV. All financial data before 1 February 2014 is dropped after the estimation of abnormal returns (described in *Section 4.1*). Thus, the dataset presented below ranges from February 2014 to June 2015 and include: financial data as defined above, computed abnormal returns and our constructed sentiment proxies.

Table III

Summary statistics – predictive regressions

This table shows the summary statistics for the underlying data for the predictive regressions (1–4). The number of non-zero sentiment scores ($SSc \neq 0$) amount to 18,326, of which 13,468 are positive and 4,858 are negative. The sum of all non-zero items ($Items \neq 0$) amount to 69,215 and is the total number of sentiment observations. The number of positive sentiment shocks (Neg_SSh = 1) amount to 590. The number of positive sentiment shocks (Pos_SSh = 1) amount to 2,410. Only one variable for abnormal return is reported (*AbnRet*) as the underlying data is the same for all. Weekday variables are excluded as they are dummies, thus statistics below do not apply. Carhart (1997) risk factors (R_{Mkt}, *SMB*, *HML*, UMD) are the same across companies (number of observations amount to 344 (R_{Mkt}) and 337 (*SMB*, *HML*, UMD)). The sample includes 286 stocks from February 2014 to June 2015.

Variable	Number of observations	Mean	Standard deviation	Minimun	Maximum
SSc	98,599	0.2281	1.5487	-45.0000	48.0000
Neg_SSh	98,599	0.0060	0.0771	0.0000	1.0000
Pos_SSh	98,599	0.0244	0.1544	0.0000	1.0000
Items	98,599	0.7020	2.6974	0.0000	141.0000
AbnRet	96,344	0.0000	0.0358	-0.5974	1.3387
R_{Mkt}	98,599	0.0008	0.0091	-0.0281	0.0306
SMB	96,344	-0.0001	0.5255	-1.5502	1.7292
HML	96,344	-0.0002	0.4077	-1.2702	1.2772
UMD	96,344	0.0002	0.5730	-2.6438	1.3925
Log_turnover	95,053	14.0202	3.0346	-0.9862	22.7111

[30] Exclusion criteria typically included in existing literature, *e.g.* Antweiler and Frank (2004) among others.

Table IV
Summary statistics – trading strategy

This table shows the summary statistics for the underlying data for the regression of sentiment-based portfolio. The sample includes 286 stocks from February 2014 to June 2015.

Variable	Number of observations	Mean	Standard deviation	Minimun	Maximum
R_{Mkt}	320	0.0008	0.0104	-0.0357	0.0605
$Port\_R_t$	350	0.0008	0.0091	-0.0281	0.0306
SMB	342	-0.0125	0.5271	-1.5502	1.7292
HML	342	-0.0149	0.4084	-1.2702	1.2772
UMD	342	0.0158	0.5733	-2.6438	1.3925

In Table III and Table IV the variables are defined as follows: *SSc* is the Sentiment Score proxy; *Pos_SSh* and *Neg_SSh* are the positive and negative Sentiment Shock proxies; *Items* is the number of sentiment observations in the corresponding day; *AbnRet* is the Abnormal Return estimated using the Carhart (1997) four-factor model; R_{Mkt} is the Excess Market Return; *SMB* is the size factor; *HML* is the book-to-market factor; *UMD* is the momentum factor; *Log_turnover* is the logarithm of the daily stock trading volumes and; $Port\_R_t$ is the Excess Portolio Return.

3.5 Discussion

When examining the effect of investor sentiment on the stock market, economists face the challenge of creating a proxy for investor sentiment. A range of different approximation methods has been used in the past (see Table I in *Section 2.1* for a list of sentiment proxies used in previous studies). We construct our novel sentiment proxies based on sentiment observations collected through data mining and the use of machine learning algorithms (provided by Modular Streams).

We believe that our method and underlying data for the construction of our proxies will add value to this field of research in the following aspects. In terms of representing the consensus sentiment for the whole market, our sentiment observations are collected from a wide range of online sources, whereas pervious research often limit the sources to one or two. For example, sources like Facebook[31], Twitter[32], Google[33] and news articles[34] are often studied in isolation. Also, our data is constructed on stock-level rather than market-level, adding to the sophistication of our analysis. Even though we are not alone in focusing on stock-level sentiment, the majority of previous studies focus on a less sophisticated market-level sentiment, *e.g.* Baker and Wurgler (2007) and Tetlock (2007) among others. However, compared to Tetlock et al. (2008), which also focus on stock-level sentiment, we extend their one-sided measure of sentiment (degrees of pessimism), by capturing both positive and

[31] Bollen et al. (2011)
[32] Karabulut (2011)
[33] Da et al. (2011)
[34] Tetlock (2007), Tetlock et al (2008)

negative sentiment in a single proxy (*SSc*) as well as separating positive and negative sentiment (*Pos_SSh* and *Neg_SSh*) in order to analyse the effects independently.

Furthermore, we concentrate our analysis on the Swedish stock market. Although Baker et al. (2012) document co-movements in investor sentiment between global and local markets, a study focusing on the Swedish stock market in isolation is the very first of its kind in this field of study. Also, even though our sample period is relatively short, ranging from February 2014 to June 2015, it provides us with the most current dataset in this field of study (see Table I in *Section 2.1* for a list of sample geographies and periods in previous studies). It can be argued that a longer sample period and broader geographical scope would be preferable in order to ensure robust results. However, we believe that due to the uniqueness of our data collection method and sentiment analysis, the underlying data will provide us with results that will add value to current research.

We want to re-emphasize the rationale behind *SSc* and *SSh*. When constructing *SSc*, as mentioned, we intend to mimic the proxy used Niederhoffer (1971). We extend the proxy used Niederhoffer (1971) by creating a more detailed (wider) range and company-specific sentiment proxy. It can be argued that by simply aggregating the sentiment observations' scores (+1 for positive, 0 for neutral and -1 for negative sentiment) per company and per day, the *SSc* also include information on posting activity volumes to some extent. In our definition of sentiment, we only address the opinion (optimism/pessimism) and not the posting activity volume. Thus, it can be argued that a standardized proxy, similar to the one used in Tetlock (2008), or a logarithmic scale can be preferred. However, we try to mitigate the effect of posting activity volume by controlling for the number of sentiment observations per day (*Items*).

As mentioned, the proxy Sentiment Shock (*SSh*), is a pair of dummy variables (*Pos_SSh* and *Neg_SSh*), taking the value of 1 if *SSc* exceed +/-3, respectively, and are otherwise equal to 0. It can be argued that by defining the variables as binary, the proxy does not account for the magnitude of the underlying shock, which is true, and therefore limits the economic interpretation of the corresponding coefficients. However, by using binary variables, it mitigates the risk of a skewed distribution distorting final results. Also, by setting the "trigger" threshold at +/-3, it reduces the noise that might occur around 0 (*SSc*). Thus, we obtain a proxy that discriminates between positive and negative sentiment and reflects sentiment shocks that "have gained some traction in the market" (>3 sentiment observations).

Even though both proxies have their individual limitations, the aggregated analysis based on the results from both proxies in combination will provide sufficient understanding in order to draw conclusions regarding our research questions.

4 METHOD

In this section, we outline the empirical methods we use to examine our research questions. We conduct five main tests using Ordinary Least Squares (OLS) regressions, in addition to the computation of abnormal returns; four predictive regressions using our computed sentiment proxies, two of which testing the relationship between investor sentiment and next-day abnormal returns, and two using the next-day excess return, testing the same relationship to ensure that firms' return covariance with priced risk factors do not drive our results (*i.e.* we avoid the joint hypothesis problem); and one regression testing our constructed risk adjusted sentiment-based trading strategy.[35] We use our compiled dataset, described in *Section 3*, as the underlying data in our regressions. All tests are conducted with daily returns in order to capture the immediate effect of investor sentiment. In addition, since our main influence is Tetlock et al. (2008), it is preferable to have comparable results in order to benchmark our analysis.

4.1 Abnormal returns

Before the main statistical analysis, we compute the Abnormal Return (*AbnRet*) for our 286 sample stocks with an estimation period for expected returns from January 2010 to June 2015. *AbnRet* will be the dependent variable in the two first main predictive regressions, described in the *Section 4.3*. We define excess return following the equations below and will refer to it as such hereon after.

$$R_{i,t} = r_{i,t} - R_{f,t-1} \ ; \ R_{Mkt,t} = r_{Mkt,t} - R_{f,t-1}$$

Extending our research methods compared to Tetlock (2008), we use the Carhart (1997) four-factor model instead of Fama-French (1993) three-factor model,

$$R_{i,t} = \alpha_i + \beta_{Mkt,i} * R_{Mkt,t} + \beta_{SMB_i} * SMB_t + \beta_{HML_i} * HML_t + \beta_{UMD_i} * UMD_t + \varepsilon_{i,t}$$

to adjust the abnormal returns for the impact of excess market returns (R_{Mkt}), size (*SMB*), book-to-market (*HML*) and momentum (*UMD*). Market returns are based on the OMX Stockholm All-share Cap GI index and the overnight STIBOR rate is used as the risk-free rate (see *Section 3*). The factor data for *SMB, HML* and *UMD* comprise US data which is converted into Swedish factor data to adjust for the FX effect on the factor returns, according to the method mentioned in Calvet et al. (2007), described in the *Section 3.1*.

---

[35] Tetlock et al. (2008)

As mentioned, we reiterate the four-factor model regression for each sample stock and predict the Expected Excess Return ($E(R_i)$) of each stock based on the estimation period, according to the following equation:

$$E(R_{i,t}) = \alpha_i + \beta_{Mkt,i} * R_{Mkt,t} + \beta_{SMB,i} * SMB_t + \beta_{HML,i} * HML_t + \beta_{UMD,i} * UMD_t$$

We then deduct the Expected Excess Return for each sample stock from the actual Excess Return for each sample stock to arrive at the Abnormal Return.

$$AbnRet_{i,t} = R_{i,t} - E(R_{i,t}) = \varepsilon_{i,t}$$

4.2 Predictive regressions

This subsection is the first of two that describes our empirical methods that aim to address our research questions. We run two sets of OLS regressions to examine the casual effect between our computed sentiment proxies (Sentiment Score (*SSc*) and the Positive and Negative Sentiment Shock (*Pos_SSh* and *Neg_SSh*)) and next-day stock returns. We use robust standard errors in our regressions.[36] All regressions are run on a dataset including 286 stocks with the sample period from February 2014 to June 2015. The first set uses the computed next-day Abnormal Return ($AbnRet_{t+1}$) as the dependent variable, while the second set uses next-day Excess Return (R_{t+1}).[37] Each set uses *SSc* in one regression and *Pos_SSh* and *Neg_SSh* in the second regression as the main independent variables. Both sets of regressions use the same range of control variables; number of sentiment observations (*Items*)[38]; day-of-the-week dummy controls (*Weekday_1*, *Weekday_2*, *Weekday_3*, *Weekday_4* and *Weekday_5*)[39]; autocorrelation in Abnormal Returns ($AbnRet_{t-1}$, $AbnRet_{t-2}$)[40]; Carhart (1993) four-factors (R_{Mkt}, *SMB*, *HML*, *UMD*) and the logarithm of the daily turnover for each stock (*Log_turnover*).[41] The regression equation is specified as:

$$RETURN_{i,t+1} = a_i + \beta_{SENT} * SENTIMENT_{i,t} + \sum \beta_{CONT} * CONTROL_t + \varepsilon_i$$

---

36 White (1980)

37 As mentioned in *Section 4* we use both abnormal returns and excess returns to ensure that firms' return covariance with priced risk factors do not drive our results (Tetlock et al. 2008), *i.e.* we avoid the joint hypothesis problem.

38 Similar to the number of news items used by Mitchell and Mulherin (1994).

39 Among others, Cross (1973), French (1980) and Gibbons and Hess (1981).

40 Tetlock et al. (2008)

41 Hong and Stein (2007)

where $RETURN_{i,\,t+1}$ represents $AbnRet_{i,\,t+1}$ for the first set of regressions and $R_{i,\,t+1}$ for the second set; $SENTIMENT_{i,\,t}$ represents either $SSc_{i,\,t}$ or $Pos\_SSh_{i,\,t}$ and $Neg\_SSh_{i,\,t}$; $CONTROL_t$ includes the control variables listed above; ε_i is the residual; α_i is the intercept.

To ensure that we have accurate statistical results, we examine if we have any issues with multicollinearity among variables in the regressions by computing a correlation matrix and running a Variance Inflation Factor (VIF) test. The results are reported in Appendix III and IV.

4.2.1 Dependent variables – returns

The main dependent variables are, as stated above, next-day Abnormal Return ($AbnRet_{t+1}$) and Excess Return (R_{t+1}). It is important to note that we regress returns at day $t+1$ against sentiment proxies and controls at day t (except for $AbnRet_t$, $AbnRet_{t-1}$, $AbnRet_{t-2}$). This allows us to disentangle the causality between investor sentiment and returns, *i.e.* does investor sentiment drive returns or do returns drive investor sentiment? By the sequential pairing of sentiment proxies and returns ($SENTIMENT_t$ paired with $RETURN_{t+1}$) we can rule out the possibility that our test results show returns driving sentiment. Also, following the methods of previous research allows us to compare our results and benchmark our analysis and conclusions.[42]

4.2.2 Independent variables – sentiment proxies

The main independent variables are, as stated above, Sentiment Score (*SSc*) and two dummy variables for positive (*Pos_SSh*) and negative sentiment shock (*Neg_SSh*). *SSc* is a discrete variable defined as the sum of the sentiment observations for each day and each company (described in detail in *Section 3.2*). The use of *SSc* is inspired by the proxy used in Neiderhoffer (1971), which is a discrete variable ranging from 1 to 7 based on the average sentiment score appointed by a panel of individuals for each news article in the data sample. Opposed to Neiderhoffer (1971), we use a non-restricted range including both negative and positive integer values. To mitigate issues with extreme values creating skewed results, we exclude observations when *SSc* exceeds +/-50.[43]

A limitation with the variable *SSc* is that one cannot differentiate the relationship between stock returns and positive and negative investor sentiment, simply because *SSc* is composed of both. In order to make this possible, addressing our second research question, we use a second proxy for sentiment, positive (*Pos_SSh*) and negative sentiment shock (*Neg_SSh*), which are binary variables taking the value of 1 if *SSc* exceeds +/-3 and otherwise

[42] Tetlock (2007) and Tetlock et al. (2008) among others.
[43] Excluded observations with an SSc value exceeding +/-50 sum to 7 observations.

taking the value 0. This allows us to determine whether positive sentiment or negative sentiment individually has a statistically significant relationship with stock returns.

In addition to providing independent results for positive and negative sentiment, the use of the second "shock-proxy" helps to ensure, or at least provide an indication of, that our results are robust. By singling out positive and negative sentiment in a binary way, one mitigates potential measurement risks (*e.g.* a lengthy forum discussion or several "re-tweets" would result in several sentiment observations (larger absolute value of *SSc*), although perhaps not representing the consensus investor sentiment accurately).

4.2.3 Control variables

We include a number of controls that are consistent throughout both sets of regressions in order to more accurately estimate the effect of investor sentiment on stock returns. The first control is for number of sentiment observations (*Items*), following Mitchell and Mulherin (1994), which concludes that the number of news announcements on Dow Jones News Wire has a positive effect on stock returns. By including this, we extract the effect that an increased/decreased amount of publicity would have from investor sentiment, regardless of positive or negative, addressing the classic saying – "All publicity is good publicity". We control for daily seasonal effects by including five dummy variables, one for each trading day (*Weekday_1, Weekday_2, Weekday_3, Weekday_4* and *Weekday_5*), following among others, Cross (1973), French (1980) and Gibbons and Hess (1981).

To control for short-term momentum, we include the Abnormal Return for days *t*, *t-1* and *t-2* ($AbnRet_t$, $AbnRet_{t-1}$, $AbnRet_{t-2}$), following Tetlock et al. (2008). This addresses the issue of autocorrelation in our dependent variables, *i.e.* returns in day *t*, *t-1* and *t-2* drives returns in day *t+1*.

Following mainstream research, we include controls for Excess Market Return (R_{Mkt}), size (*SMB*), book-to-market (*HML*) and momentum (*UMD*) in order to extract market wide effects on stock specific returns.[44] Also, we control for trading volume by including the logarithm of daily stock turnover, following Hong and Stein (2007) among others.[45]

4.2.4 Discussion

Main tests in Tetlock (2007) focus on a proxy for sentiment constructed by the standardized fraction of negative words in news stories and finds a significant negative relationship between the sentiment proxy and Dow Jones returns. However, it is mentioned that in unreported tests, similar results are found for a proxy including both negative and positive words as a proxy for sentiment. This is later supported by Tetlock et al. (2008),

---

[44] Fama and French (1993) and Carhart (1997)
[45] Da et al. (2011) and Tetlock et al. (2008)

where unreported tests show significant results for both positive and negative words (proxy for positive and negative sentiment), although results for negative words indicate a stronger effect. Thus we expect both *Pos_SSh* and *Neg_SSh* (and *SSc*) to show significant results while the loadings on *Neg_SSh* is expected to be larger in absolute magnitude.

Although our methods are similar to the ones used in several previous papers, we extend the scope of the research in several ways. The use of more comprehensive sentiment proxies; including positive sentiment (positive values of *SSc* and *Pos_SSh*) is an extension of the research of Tetlock (2007) and Tetlock et al. (2008), which only officially conduct tests on a proxy for negative sentiment. The use of both the discrete and binary sentiment proxies further ensures robust results, and is not done by Tetlock et al. (2008). Other distinguishing factors with our thesis relating to the dataset are discussed in *Section 3.4*.

We make an implicit assumption regarding our regression estimates: we assume the estimates will remain stable during the sample period as we do not use rolling-window estimation. Given our short sample period and daily frequency of data, we argue that by applying a rolling-window method, it would cause more issues rather than improving our regression results. Also, this means that we conduct in-sample-estimation, which is not entirely realistic since information for day $t+1$ is not known in day t. However, we believe that it will not invalidate our final results.

4.3 Sentiment-based trading strategy

This subsection is the second and last that describes our empirical methods. If time-varying investor sentiment predicts next-day stock returns, investors should be able to exploit the cross-sectional trading opportunities. Signals indicated by *SSc* should thus be a useful tool for timing the market. We construct an equally weighted long-short portfolio based on the sentiment proxy *SSc* and run an OLS regression based on the Carhart (1997) four-factor model using our FX adjusted factors (R_{Mkt}, *SMB*, *HML*, *UMD*) with robust standard errors.[46] The excess return of the long-short portfolio ($Port\_R_t$) is the dependent variable. We intend to in part test if our sentiment proxy can be used as a trading signal, addressing our third research question, but also to provide additional assurance that our previous regressions provide robust results, by limiting the effect of idiosyncratic risk on the results by constructing portfolios of stocks.

The results from the regression can then be used to compute the theoretical annualized risk-adjusted return of the portfolio following the trading strategy (described in *Section 4.4.2*. Results are reported in *Section 5.4*).

[46] White (1980)

4.3.1 Portfolio construction

We construct the equally weighted long-short portfolio with *SSc* as the governing variable. Each day, the 286 stocks in our sample from February 2014 to June 2015 are ranked after their individual *SSc* from 0 to 286, where 0 is the lowest and 286 is the highest. We go long in the five stocks with highest rank and short the five stocks with lowest rank at market closing each day (day *t*). We hold the positions for one full trading day, rebalancing our portfolio at the closing of the following trading day (day *t+1*). This is done every day for our sample period. The constructed portfolio will yield returns each day throughout our sample period according to the following equation,

$$Port\_r_t = \frac{1}{5} * \sum_{i=1}^{5} r_{i,t}^{Pos} - \frac{1}{5} * \sum_{i=1}^{5} r_{i,t}^{Neg}$$

where $r_{i,t}^{Pos}$ represents the returns of highest ranked stocks each day and $r_{i,t}^{Neg}$ represents the returns for the lowest ranked stocks each day. The risk-free rate is then deducted from the portfolio returns in order to get the Excess Portfolio Return ($Port\_R_t$),

$$Port\_R_t = Port\_r_t - R_{f,t-1}$$

which is used as the dependent variable in the regression. In days when there are insufficient positive or negative sentiment observations in order to construct our portfolio (less than five positive and five negative observations), we take no position in the market. Neutral sentiment observations are never included in the portfolio. Following the majority of finance papers, we assume an efficient market without trading costs, capital gains tax, etc.

We base our trading strategy on a number of decisions regarding portfolio construction and holding period. We decide to include the top five and bottom five stocks in our portfolio, ten positions in total. While the praxis in finance research using portfolio construction methods is to use quartiles or deciles for defining the portfolio inclusion range, we have to adjust the method to the availability of data in our sentiment dataset. We note that the portfolio size, including only ten positions, is relatively small, and would reduce the desired effect of ridding the idiosyncratic risk. However, we test other portfolio sizes, increasing the total number of held stocks to twelve, fourteen and sixteen, aiming to increase the quality of our results by diversifying the idiosyncratic risk. However, we do not find that this improves the quality of our results.[47]

[47] With the increasing number of stocks included, the number of observations decrease. Also, as the mean of the *SSc* lies around 0.23 with a standard deviation of 1.55, most observations are close to the mean. Thus, it is not

The holding period is also a factor that has an important effect on our tests, even more so than the portfolio size. By extending the holding period, the underlying question that the regression aims to shed light upon changes, *e.g.* by increasing the holding period to 2 or 3 days, one can investigate drifts in stock prices following a sentiment observation. For this reason and in order to report comparable results to our main inspiration paper (Tetlock et al. (2008)), we choose a one-day holding period for the sentiment-based portfolio.

4.3.2 Portfolio regression

To better examine the performance of the sentiment-based portfolio, we run an OLS regression against the Carhart (1997) four-factor model (R_{Mkt}, *SMB*, *HML* and *UMD*) with Excess Portfolio Returns ($Port\_R_t$) as the dependent variable. The regression is specified as follows:

$$Port\_R_t = \alpha + \beta_{Mkt} * R_{Mkt,t} + \beta_{SMB} * SMB_t + \beta_{HML} * HML_t + \beta_{UMD} * UMD_t + \varepsilon$$

where α, the intercept, is our variable of focus, representing the daily risk-adjusted excess portfolio return of the trading strategy.[48]

We expect to find an intercept that indicates substantial risk-adjusted portfolio returns. Previous research has found similar results; Tetlock et al. (2008) among others, document consistent daily returns of 10.13 bps on average over their sample period from 1980 to 2004. By annualizing the return, Tetlock et al. (2008) document an average risk-adjusted return of 21.1 percent per year before trading costs.

---

certain that increasing the number of stocks improves the desired results as it might be one or two stocks with a Sentiment Score above (below) +2 or +3 (-2 or -3) while the majority of non-zero *SSc* is +1 (-1), *i.e.* the assigned rank of 0 to 286 becomes arbitrary when the majority of stocks have the same (small) *SSc*.

[48] Annualized returns are computed as (1 + α) ^ 260 – 1. 260 is the normal amount of total trading days in Nasdaq Stockholm

5 RESULTS AND ANALYSIS

This section reports our main results and analysis of our predictive regressions as well as our trading strategy. Further results regarding correlation matrices and VIF tests are presented in the appendix (*Section 8*).

5.1 Predictive regressions

We address the first and second research questions – if next-day stock returns can be predicted by investor sentiment and if the predictive characteristics differ when discriminating between positive and negative sentiment – by running two sets of predictive regressions to estimate the effect of *SSc* and *SSh* on next-day returns measures ($AbnRet_{t+1}$ and R_{t+1}). Regression coefficients, significance levels and R^2-values are reported in Table V.

We document that all three of our constructed sentiment proxies (*SSc*, *Pos_SSh* and *Neg_SSh*) contribute to predicting next-day stock returns after adding controls (listed in Table V). *SSc* is positively related to both next-day $AbnRet_{t+1}$ and R_{t+1}, *i.e.* the more positive investor sentiment relating to a specific stock in day *t*, the higher the stock return in day *t+1*. One unit increase in *SSc* predicts an increase of 4 basis points in of R_{t+1} and 3 basis points increase in $AbnRet_{t+1}$. The average R^2 equals 0.018. In line with *SSc* and our expectations, *Pos_SSh* and *Neg_SSh* predict positive and negative next-day stock returns ($AbnRet_{t+1}$ and R_{t+1}), respectively. All coefficients for our sentiment proxies are statistically significant throughout all four predictive regressions with a p-value below one percent. Thus, we can reject the null hypothesis that the coefficients equal to zero. The results we document in the predictive regressions contradict the findings in Roll (1988), which claims that news articles cannot affect stock prices. However, our results confirm the findings in Tetlock (2007) and Tetlock et al. (2008).

When discriminating between positive and negative investor sentiment, using *Pos_SSh* and *Neg_SSh*, we document that negative sentiment has a larger absolute effect on next-day returns than positive sentiment. This is consistent with the findings of Chan (2003) and Tetlock et al. (2008). For both of the dependent variables (R_{t+1} and $AbnRet_{t+1}$), the economic significance of negative sentiment is approximately 55 percent larger than for positive sentiment. One could argue that the difference originates from the construction of *SSh*, as it does not take into account the magnitude of *SSc* exceeding the cut-off point (*i.e. SSc* = 4 and *SSc* = 10 are treated the same).[49] However, when examining the underlying data, we find that the mean of positive values in *SSc* is larger than the absolute mean of negative

[49] Different threshold values are tested in unreported tests, ranging from 3 to 8. Results show a decreasing statistical significance of the related coefficients as the threshold value increases and the number of “triggering” observations decrease.

values in *SSc*. In other words, we find that investors have distinct sensitivities to positive and negative sentiment and that, investors are more reactive to negative sentiment, *i.e.* investor decisions are affected to a greater extent by the risk of losing money than the chance of earning money. This is in line with the theory *Loss Aversion* within behavioural finance.

Table V

Predicting returns using sentiment proxies

This table shows the relationship between the sentiment proxies (*SSc*, *Neg_SSh* and *Pos_SSh*) and firms' abnormal ($AbnRet_{t+1}$) and excess returns (R_{t+1}) on the following trading day for the time period February 2014 to June 2015, through four regressions. The results are displayed below. The dependent variable $AbnRet_{t+1}$ is derived from the Carhart (1997) four-factor model with a benchmark period for expected returns ranging from January 2010 to June 2015. The main independent variables are sentiment proxies based on machine learning and sentiment analysis on online sources and differ between regressions. The proxy *SSc* ("Sentiment Score"), which is the sum of each day's sentiment observations per company; the proxy *Neg_SSh* ("Negative Sentiment Shock") and *Pos_SSh* ("Positive Sentiment Shock"), are dummy variables that equal to 1 if *SSc* for each day and company is smaller (larger) than -3 (+3), respectively, and equal to 0 otherwise. Each regression also includes a number of controls as displayed below. We use robust standard errors. The significance level is derived from the p-value and displayed for each coefficient with "*".

Dependent variable	R_{t+1}	R_{t+1}	$AbnRet_{t+1}$	$AbnRet_{t+1}$
Sentiment proxy	*SSc*	*Neg_SSh* *Pos_SSh*	*SSc*	*Neg_SSh* *Pos_SSh*
SSc	0.0004 ***	*n.a.*	0.0003 ***	*n.a.*
Neg_SSh	*n.a.*	-0.0044 ***	*n.a.*	-0.0040 ***
Pos_SSh	*n.a.*	0.0028 ***	*n.a.*	0.0026 ***
Items	0.0000	0.0000	-0.0001	0.0000
Weekday_1	-0.0006	-0.0006	*(omitted)*	*(omitted)*
Weekday_2	0.0003	0.0003	0.0011 ***	0.0011 ***
Weekday_3	0.0007	0.0007 *	0.0009 **	0.0009 **
Weekday_4	0.0017 ***	0.0017 ***	0.0011 ***	0.0011 ***
Weekday_5	*(omitted)*	*(omitted)*	-0.0004	-0.0004
$AbnRet_t$	-0.1222 ***	-0.1219 ***	-0.1215 ***	-0.1212 ***
$AbnRet_{t-1}$	-0.0370 ***	-0.0369 ***	-0.0388 ***	-0.0387 ***
$AbnRet_{t-2}$	-0.0265 *	-0.0265 *	-0.0259 *	-0.0259 *
R_{Mkt}	0.0761 ***	0.0764 ***	0.1148 ***	0.1150 ***
SMB	0.1581 ***	0.1578 ***	0.0712 ***	0.0709 ***
HML	-0.1180 ***	-0.1184 ***	-0.0711 *	-0.0713 *
UMD	0.0430 *	0.0427 *	-0.0618 **	-0.0620 **
Log_turnover	0.0000	0.0000	0.0000	0.0000
Intercept	0.0000	0.0000	-0.0005	-0.0005
Number of observations	88469	88469	86015	86015
R^2	0.0182	0.0182	0.0179	0.0179

* 0.1 > p > 0.05

** 0.05 > p > 0.01

*** p < 0.01

By comparing the four regressions reported in Table V, we can assert that our sentiment proxies robustly predict next-day stock returns. The magnitudes and signs of the coefficients as well as the statistical significance are very similar in all four regressions. Also, as there is no significant difference between the results when using R_{t+1} and $AbnRet_{t+1}$ as the dependent variable, our results are not driven by firms' return covariance with priced risk factors. Further, we document statistical significant coefficients corresponding to R_{Mkt}, *SMB*, *HML*, *UMD*, $AbnRet_t$, $AbnRet_{t-1}$, $AbnRet_{t-2}$, some of the weekday dummies and *SSc*, *Pos_SSh* and *Neg_SSh* across all four predictive regressions. Coefficients for *Log_turnover* and *Items* are not statistically significant. The controls $AbnRet_t$, $AbnRet_{t-1}$, $AbnRet_{t-2}$, suggest that we observe short-term return reversals in our sample, as they have negative coefficients. As can be expected, the statistical significance and magnitude of the coefficients for *AbnRet* decline the further from day $t+1$ the observation is.

We ensure that our regressions are not influenced by problematic multicollinearity by running a Variance Inflation Factor (VIF) test and computing a correlation matrix with our independent variables (reported in Appendix III and Appendix IV). We document that no correlation between independent variables exceed 0.5 and no VIF value exceeds 2.0. This confirms that we do not have an issue with multicollinearity in our regression results.

5.1.1 Discussion

By using similar empirical methods to existing research but other sentiment proxies we can easily benchmark our results and form a discussion on the differences and similarities we observe. Tetlock (2007) and Tetlock et al. (2008) conclude that their sentiment proxy, based on the fraction of negative words in texts, predicts next-day stock returns. They regress their proxy on raw stock returns and abnormal returns and control for other commonly used factors. We document similar results to Tetlock et al. (2008), thus we argue that our sentiment proxies capture investor sentiment in a similar way as the proxy used in Tetlock (2007) and Tetlock et al. (2008). Additionally, in unreported tests, Tetlock et al. (2008) claim to find a stronger effect when using a proxy for negative sentiment compared to a proxy for positive sentiment. We find the same phenomena when examining the differences in *Pos_SSh* and *Neg_SSh*; in extension to Tetlock et al. (2008) we quantify the difference to be 55 percent. However, as *SSh* does not account for that the underlying sentiment data is biased toward positive sentiment, we cannot conclude that the magnitude of investors' reaction to negative sentiment is exactly 55 percent larger than their reaction to positive sentiment.

Another difference compared to Tetlock et al. (2008) is the sample geography. Tetlock et al. (2008), and the majority of research covering investor sentiment, use samples based on US equities. We extend current research and the understanding of investor sentiment by basing our sample on Swedish equities.[50] As we document similar results to Tetlock et al. (2008), we argue that investor sentiment robustly predicts stock returns across geographies. This is also argued by Baker et al. (2012) which uses multiple samples from Canada, France, Germany, Japan, the United Kingdom, and the United States. They too document significant

[50] Other papers using the US market can be viewed in Table I in *Section 2.1*

results suggesting a positive relationship between investor sentiment and stock returns. Further, it can be concluded that the persistence across markets of the examined effect of investor sentiment is not due to faults in the empirical methods or sentiment proxies we use. Compared to Baker et al. (2012), both our empirical methods as well as our sentiment proxies differ significantly.

Investor sentiment is also found to be persistent across different time periods. As can be seen in Table I in *Section 2.1*, Tetlock (2007), Tetlock et el. (2008), Baker and Wurgler (2006 and 2007), Baker et al. (2012), Livnat and Petrovits (2009) and Antweiler and Frank (2004), collectively cover the time period from 1926 to 2005 in their data samples. All of which also document confirming evidence that investor sentiment helps to predict stock returns (or market returns). We extend this by using a sample including 2014 and 2015. Thus we argue that the examined effect of investor sentiment on stock returns is still present today and has not been diminished by recent market developments (*e.g.* advances in high frequency trading).

An aspect that further differentiates our research is the use of sources for our sentiment proxies. It can be argued that by including Twitter, Facebook, chat forums etc. as sources, it adds unwanted noise to our proxies. However, we infer that this captures the sentiment of retail investors to a certain extent. This is previously done by for example Antweiler and Frank (2004) and Da et al. (2011), which bases their analysis on the disagreements on stock message boards and search frequencies for specific stocks on Google, respectively. Both document significant findings that are broadly consistent with noise trader models – capturing systematic activities of retail investors affecting stock returns. This is supported by Kumar and Lee (2006), which finds that stocks that are generally preferred by retail investors show a stronger relationship to proxies of retail investor sentiment. However, even though noise trader models show significant results, we argue that they fail to capture the consensus sentiment for the market as a whole. Thus the inclusion of a broader range of sources when constructing our sentiment proxies, should give us a better understanding, and more robust results, of how investor sentiment affects stock returns.

Resulting from our choice of empirical methods, limited to focus on next-day stock returns, we cannot conclude whether investors underreact or overreact to sentiment signals. Previous studies extend their focus time period to investigate week-long or month-long returns following sentiment events. By doing so, authors increase the understanding of investor behaviour and can not only comment on if sentiment has a significant effect, but also the characteristic of the effect. Neiderhoffer (1971) documents return reversals following extremely bad news the subsequent 2-5 days suggesting the market overreacts to bad news. This is later confirmed by Baker et al. (2012) among others, suggesting that assets prices deviate from their intrinsic value in times of high/low sentiment and subsequently revert back to fundamentals. Livnat and Petrovits (2009), on the other hand, study the PEAD phenomenon in the context of underreaction and find that the examined effect is amplified when sorting on sentiment prior to the earnings announcement. Our results can be interpreted both in the context of overreaction and underrecation. The significant coefficients for our sentiment proxies indicate that sentiment information is not perfectly reflected in prices in day *t* suggesting either that there is a delayed market response which is reflected in next-day returns or the market fully prices the information in day *t*, but overreacts in day *t+1*, explaining the returns.

This might seem unrelated to the focus of our research questions – whether investor sentiment can predict next-day stock returns. However, Tetlock (2007) argues that by examining a prolonged time period after a sentiment event, one can distinguish between the effect of new information contained in news articles and the effect of investor sentiment. He means, "*if the column contained new information about fundamentals, there could be an initial decline in returns, but this would not be followed by a complete return reversal.*" Even though this is true, we argue that our results and analysis examine the effect of investor sentiment to a sufficiently extent, where we can draw robust conclusions regarding our research questions.

5.2 Sentiment-based trading portfolio

The return predictability of investor sentiment we find in the predictive regressions suggests the possibility that a sentiment-based trading strategy could earn substantial risk-adjusted returns. We examine the possibility by constructing a simple long-short portfolio based on rankings derived from *SSc*. The performance is adjusted for common risk factors (Carhart (1997)) and results show substantial risk-adjusted returns. Regression results are reported in Table VI. The correlation matrix and VIF test is reported in Appendix V and Appendix VI.

The variable of focus in the below regression results is the intercept. It portrays the daily risk-adjusted return for our trading strategy; we earn 8.7 basis points from February 2014 to June 2015. By annualizing the daily return, we earn a theoretical return of 25.4 percent per year. The return is under the assumption of perfect financial markets without trading costs or capital gains taxes, etc. This is similar in magnitude to the return documented in Tetlock et al. (2008) – reporting an average risk-adjusted return of 21.1 percent per year before trading costs. The factor loadings for R_{Mkt}, *SMB*, *HML* and *UMD* are reported in Table VI. Note that the number of observations amounts to 312, which is the number of days that our trading strategy is "triggered" and we take positions in the market.

As can be seen in Table VI, the intercept (implying our daily return) is not statistically significant. In other words, we cannot reject the null hypothesis that it is different from zero, *i.e.* our results can be driven by chance. Moreover, the R^2 for our regression specification is 0.011, implying that the majority of the variance of in the portfolio return cannot be explained by Carhart's (1997) four risk factors. The overall low statistical significance of our regression results can be due to the few number of observations in our sample, which is a function of our short sample period and portfolio construction criteria.

Table VI

Risk-adjusted sentiment trading strategy returns

This table shows the daily risk-adjusted excess return (*Intercept*) from a sentiment-based long-short trading strategy for the time period from February 2014 to June 2015. The equal-weighted portfolio is constructed at the close of each trading day by going long in the top 5 ranked stocks and going short in the bottom 5 ranked stocks based on *SSc* at day *t*. We hold the portfolio for one full trading day and rebalance it at the end of the next trading day (*t+1*). We take no position in days in which a full portfolio (5+5 stock) cannot be assembled due to lack of sentiment observations. The regression uses the Carhart (1997) four-factor model to control for the impact of market returns (R_{Mkt}), size (*SMB*), book-to-market (*HML*) and momentum (*UMD*). The results are displayed below. We use robust standard errors. The significance level is derived from the p-value and displayed for each coefficient with "*".

Dependent variable	*R_Port*
R_{Mkt}	0.09614 [(1)]
SMB	-0.00059
HML	-0.00048
UMD	0.00067
Intercept	0.00087
Number of observations	312
R^2	0.0110

* 0.1 > p > 0.05

** 0.05 > p > 0.01

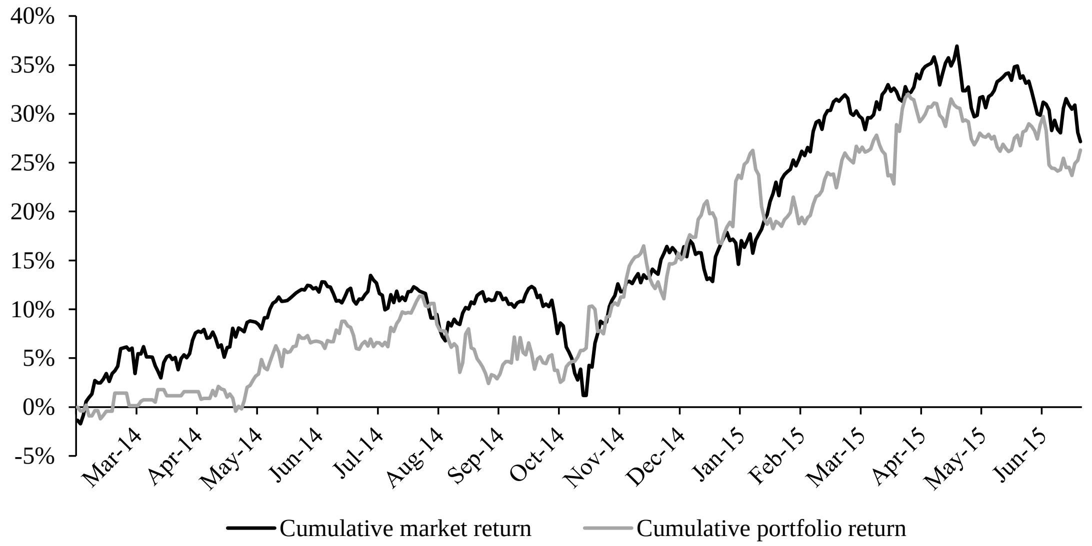
*** p < 0.01

[(1)] no coefficient show p < 0.1

We graph the performance of our trading portfolio and compare it to the cumulative market return in Graph I. We find that our portfolio return amounts to 26.3 percent for the period February 2014 to June 2015. By annualizing this, we would earn a return of 21.4 percent, which is slightly below the theoretical risk-adjusted return indicated by our regression results. Compared to the market return of 27.2 percent, the portfolio slightly underperforms the market for this specific period. Two important notes to be considered are: firstly, the portfolio performance ignores trading costs, which would quickly erode the performance as the portfolio is rebalanced every day; secondly, our portfolio is a zero cost position, while the market index is not, suggesting that the comparison with the market is somewhat misrepresentative. However, it is still interesting to observe the differences in movement compared to the market. In October 2014 the market experiences a severe drop in performance while our portfolio seems to avoid this and instead show positive returns. The same can be observed in December 2014. Further, in January 2015 the market shows a flat performance while our portfolio shows significant returns, followed by a reversal. In March

2015, the market shows slight positive returns while our portfolio earns negative returns, followed by a reversal. Nevertheless, our portfolio seems to follow the market fairly close across the period as a whole.

Graph I
Cumulative market and sentiment-based portfolio return



5.2.1 Discussion

Even though the regression results do not show any statistical significance, by analysing the signs corresponding to the factor loadings, we can extend our understanding of how our trading strategy executes. We observe a negative coefficient for *SMB*, which can be interpreted as our trading strategy tends to go long in large stocks and short small stocks. It can be argued that this is a results of our data quality assurance process, where we exclude many of the small stocks in our sample due to the lacking quality of the data relating to those stocks (described in *Section 3.3*). It can also be derived from the fact that we use factor data based on US equities (described in *Section 3.1*). Because American equities are in general larger that Swedish equities, the Swedish large cap stocks in our portfolio might show similar characteristics as American small cap stocks. The observed loading on *HML* is negative, which, following the same reasoning as for the loading on *SMB*, can be interpreted as our strategy tends to go long in growth stocks and short value stocks. Finally, the loading on *UMD* is positive, which implies that our strategy tends to go long in winner stocks and short loser stocks.

We argue that our factor loadings confirm previous research to a certain extent and capture the behaviour of retail investors. The coefficient signs match those reported in Tetlock et al. (2008), which is expected as we follow the same empirical method, even though our research is conducted on a dataset comprised of Swedish equities. Baker and Wurgler (2007) sort stocks based on factor characteristics and find that growth stocks of relative small size is affected to a greater extent by investor sentiment than value stocks of a larger size. Even though this is not directly comparable to our findings, our strategy tends to confirm the former by longing growth stocks while it opposes the latter by shorting small stocks. The positive loading on *UMD* can be related to the positive feedback theory put forth by De Long et al. (1990b), where positive (negative) sentiment is a result of preceding positive (negative) returns. It can be argued that the loadings, which is in extension a function of *SSc*, capture the behaviour of retail investors, who are most likely to be affected by news articles and other online content. Intuitively, larger companies that have had a significant performance (positive or negative) are more likely to be mentioned in the news and subsequently discussed on stock message boards, chat forums etc., thus causing an increased reading in *SSc* and therefore more likely to be included in our portfolio.

6 CONCLUSIONS

The paper examines the effect of investor sentiment on next-day stock returns. We approximate investor sentiment through the use of computer science techniques including data mining and machine learning. Our empirical methods consist of OLS regression analysis, conducted on our sentiment proxies (Sentiment Score (*SSc*) and Sentiment Shock (*SSh*)) and a number of controls commonly used in previous research. We concentrate our analysis on 286 stocks trading on Nasdaq Stockholm and Nasdaq First North during the time period from February 2014 to June 2015.

Our main results show that investor sentiment predicts next-day stock returns – positive (negative) sentiment predicts higher (lower) return in the subsequent trading day. By using similar empirical methods and documenting similar results to previous research (Tetlock et al. (2008)), we conclude that our sentiment proxies robustly captures investor sentiment. We extend previous research by using unique sentiment approximation methods, sample geography and sample time period. Our results broadly confirm results of previous research (discussed in *Section 5.1*), thus we conclude that investor sentiment's effect on subsequent stock returns is persistent across geographies, time periods and robust to different empirical methods.

We find that the characteristics of the relationship between investor sentiment and next-day stock returns differ when discriminating between positive and negative sentiment. We find that negative investor sentiment has a larger absolute economic effect on next-day stock returns than positive investor sentiment, which confirms findings of Chan (2003), Tetlock (2007) and Tetlock et al. (2008) and the *Loss Aversion* theory within behavioural finance.

Following the confirming evidence of return predictability, we construct an equally weighted long-short sentiment-based trading strategy and model the theoretical return of the strategy. We document a risk-adjusted return of 25.4 percent per year when ignoring trading costs. This is in line with the magnitude documented by Tetlock et al. (2008). However, our results are not statistically significant, thus we cannot conclude that a trading strategy constructed on a sentiment-based algorithm will yield positive returns. However, we test the trading strategy for our sample period and find that the actual return of the portfolio amounts to 26.3 percent, ignoring trading costs. Even though our results are statistically insignificant, collectively they show an indication of that sentiment-based trading strategies could earn substantial positive returns.

The documented results collectively extend current research and the understanding of how investor sentiment affects asset prices. Our findings broadly confirm findings in previous literature, even though some studies report contradicting results. The empirical methods used in this study are influenced by Tetlock et al. (2008) and we do not have reason to question the robustness of our results due to faults in the empirical methods. Our sample period is relatively short and our sample of stocks is relatively small, compared to some previous research. This is in part a function of our data quality control criteria, where we exclude a large amount of stocks from our sample and a result of the sentiment dataset obtained from Modular Streams, which is only collected starting from February 2014. We conclude that the overall quality and robustness of the results from our predictive regressions is sufficient, however the results from our trading strategy would benefit from conducting back-tests on a larger sample and longer sample period.

6.1 Future research

We encourage a number of extensions in future research efforts. Firstly, we promote the use of sentiment proxies constructed from the use of data mining and machine learning algorithms. These methods become increasingly sophisticated each year and can be argued to be superior to other methods, not only in the quantity of data they can efficiently collect, but also the accuracy of the sentiment captured. Proxies measuring search volumes, message board posting volumes, price development, fraction of negative words in news articles, etc. might be equal in regards to the quantity of data they can efficiently collect, however they fail to accurately capture actual opinions of investors.

Secondly, we suggest that future research efforts extend their geographical scope to include markets other than the United States. Baker et al. (2012) do this to an extent, covering some of the European markets, Japan and North America including the US and Canada. However, as more financial markets become more developed, it facilitates the testing in new markets. Also, studying investor sentiment in less developed financial markets, if possible, is interesting as developed markets share many similar characteristics, while less developed markets might behave differently.

Lastly, following previous research, we propose studying investor sentiment during longer event windows in connection with discriminating features. By doing this one gains a deeper understanding of investor's behaviour in connection with sentiment events. As mentioned, we fail to distinguish between over- and underreaction as our event window is limited to next-day returns. Even though over- and underreaction has been studied in previous literature, it has never been studied in combination with for example separating negative and positive sentiment.[51] Other features might include geography, sample period, sorting on size, value, momentum, source of sentiment proxy (*e.g.* Twitter, news articles, message boards).

[51] For example Niederhoffer (1971), Da et al. (2011)

7 REFERENCES

7.1 Academic sources

Abarbanell, Jeffery S., and Victor L. Bernard, 1992, Tests of Analysts' Overreaction / Underreaction to Earnings Information as an Explanation for Anomalous Stock Price Behavior, *Journal of Finance* 47, 1181-1207.

Antweiler, Werner, and Murray Z. Frank, 2004, Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards, *Journal of Finance* 59, 1259-1294.

Bachelier, Louis, 1900, Théorie de la Spéculation, *Annales scientifiques de l'École Normale Supérieure* 3, 21-86.

Bagnoli, Mark, Messod D. Beneish, and Susan G. Watts, 1999, Whisper Forecasts of Quarterly Earnings Per Share, *Journal of Accounting and Economics* 28, 27-50.

Baker, Malcom, and Jeffrey Wurgler, 2006, Investor Sentiment and the Cross-Section of Stock Returns, *Journal of Finance* 61, 1645-1680.

Baker, Malcom, and Jeffrey Wurgler, 2007, Investor Sentiment in the Stock Market, *Journal of Economic Perspectives* 21, 129-151.

Baker, Malcolm, Jeffrey Wurgler, and Yu Yuan, 2012, Global, Local, and Contagious Investor Sentiment, *Journal of Financial Economics*, 104, 272–285.

Ball, Ray, and Philip Brown, 1968, An Empirical Evaluation of Accounting Income Numbers, *Journal of Accounting Research* 6, 59-178.

Barberis, Nicholas, Andrei Shleifer, and Robert Vishny, 1998, A Model of Investor Sentiment, *Journal of Financial Economics* 49, 307-343.

Bernard, Victor L., and J. Thomas, 1989, Post-Earnings Announcement Drift: Delayed Price Response or Risk Premium?, *Journal of Accounting Research* 27, 1-48.

Bird, Ron, Daniel F. S. Choi, and Danny Yeung, 2014, Market Uncertainty, Market Sentiment, and the Post-Earnings Announcement Drift, *Review of Quantitative Finance and Accounting* 43, 45-73.

Bird, Ron, and Danny Yeung, 2012, How Do Investors React to Uncertainty?, *Pacific-Basin Finance Journal* 20, 310-327.

Bollen, J., H. Mao, and X. Zeng, 2011, Twitter Mood Predicts the Stock Market, *Journal of Computational Science*, 2, 1-8.

Brown, Gregory W., and Michael T. Cliff, 2004, Investor Sentiment and the Near-Term Stock Market, *Journal of Empirical Finance* 11, 1-27.

Brown, Gregory W., and Michael T. Cliff, 2005, Investor Sentiment and Asset Valuation, *Journal of Business* 78, 405-440.

Calvet, Laurent E., John Y. Campbell, and Paolo Sodini, 2007, Fight or Flight? Portfolio Rebalancing by Individual Investors, Working paper, HEC Paris, Harvard University and Stockholm School of Economics.

Carhart, M. Mark, 1997, On Persistence in Mutual Fund Performance, *Journal of Finance* 52, 57–82.

Cutler, David M., James M. Poterba, and Lawrence H. Summers, 1989, What Moves Stock Prices?, *Journal of Portfolio Management* 15, 4-12.

Chan, Louis K. C., Narasimhan Jegadeesh, and Josef Lakonishok, 1999, The Profitability of Momentum Strategies, *Financial Analysts Journal* 55, 80-90.

Chan, Wesley S., 2003, Stock Price Reaction to News and No-News: Drift and Reversal After Headlines, *Journal of Financial Economics* 70, 223-260.

Chopra, Navin, Josef Lakonishok, and Jay R. Ritter, 1992, Measuring Abnormal Performance: Do Stocks Overreact?, *Journal of Financial Economics* 31, 235-268.

Cohen, Lauren, and Andrea Frazzini, 2008, Economic Links and Predictable Returns, *Journal of Finance* 63, 1977-2011.

Da, Zhi, Joseph Engelberg, and Pengjie Gao, 2011, In Search of Attention, *Journal of Finance* 66, 1461-1499.

Daniel, Kent, David Hirshleifer, and Avanidhar Subrahmanyam, 1998, Investor Psychology and Security Market Under- and Overreactions, *Journal of Finance* 53, 1839-1885.

De Bondt, Werner F. M., and Richard Thaler, 1985, Does the Stock Market Overreact?, *Journal of Finance* 40, 793-805.

De Bondt, Werner F. M., and Richard Thaler, 1987, Further Evidence On Investor Overreaction and Stock Market Seasonality, *Journal of Finance* 42, 557-581.

De Long, J.B., Shleifer, A., Summers, L., Waldmann, R., 1990b, Positive Feedback Investment Strategies and Destabilizing Rational Speculation, *Journal of Finance* 45, 375-395.

Fama, Eugene F., 1965, The Behaviour of Stock-Market Prices, *Journal of Business* 38, 34-105.

Fama, Eugene F., 1970, Efficient Capital Markets: A Review of Theory and Empirical Work, *Journal of Finance* 25, 383-417.

Fama, Eugene F., and Kenneth R. French, 1992, The Cross-Section of Expected Stock Returns, *The Journal of Finance* 47, 427-465.

Fama, Eugene F., Lawrence Fisher, Michael C. Jensen, and Richard Roll, 1969, The Adjustment of Stock Prices to New Information, *International Economic Review* 10, 1-21.

Fama, Eugene F., and Kenneth R. French, 1996, Multifactor Explanations of Assets Pricing Anomalies, *The Journal of Finance* 51, 55-84.

Gibbons, Michael R., and Hess Patrick, 1981, Day of the Week Effects and Assets Returns, *Journal of Business* Vol. 54, No.4, pp. 579-596.

Grinblatt, Mark, and Sheridan Titman, 1989, Mutual Fund Performance: An Analysis of Quarterly Portfolio Holdings, *Journal of Business* 62, 394-415.

Grinblatt, Mark, and Sheridan Titman, 1993, Performance Measurement without Benchmarks: An Examination of Mutual Fund Returns, *Journal of Business* 66, 47-68.

Hong, Harrison, and Jeremy C. Stein, 2007, Disagreement and the Stock Market, *Journal of Economic Perspectives* 21, 109-128.

Jegadeesh, Narasimhan, and Sheridan Titman, 1993, Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency, *Journal of Finance* 48, 65-91.

Jegadeesh, Narasimhan, and Sheridan Titman, 2001, Profitability of Momentum Strategies: An Evaluation of Alternative Explanations, *Journal of Finance* 56, 699-720.

Jegadeesh, Narasimhan, 1990, Evidence of Predictable Behavior of Security Returns, *Journal of Finance* 45, 881-898.

Jensen, Micheal C., and George A. Benington, 1970, Random Walks and Technical Theories: Some Additional Evidence, *Journal of Finance* 25, 469-482.

Joseph, Kissan, M. Babajide Wintoki, and Zelin Zhang, 2011, Forecasting Abnormal Stock Returns and Trading Volume Using Investor Sentiment: Evidence from Online Search, *International Journal of Forecasting* 27, 1116-1127.

Karabulut, Y., 2011, Can Facebook Predict Stock Market Activity? Working paper, Goethe University Frankfurt.

Kumar, A., and C. Lee, 2006, Retail Investor Sentiment and Return Comovements, *Journal of Finance* 61, 2451–86.

Lakonishok, Josef, Andrei Shleifer, and Robert W. Vishny, 1994, Contrarian Investment, Extrapolation, and Risk, *Journal of Finance* 49, 1541-1578.

Lehmann, Bruce, 1990, Fads, Martingales and Market Efficiency, *Quarterly Journal of Economics* 105, 1-28.

Lemmon, M., and E. Portniaguina. 2006, Consumer Confidence and Asset Prices: Some Empirical Evidence, *The Review of Financial Studies* 19:1499–529.

Levy, Robert A., 1967, Relative Strength as a Criterion for Investment Selection, *Journal of Finance* 22, 595-610.

Livnat, Joshua, Christine Petrovits, 2009, Investor Sentiment, Post-Earnings Announcement Drift, and Accruals, Working Paper, Stern School of Business.

Lo, Andrew W., and A. Craig Mackinlay, 1990, When Are Contrarian Profits Due To Stock Market Overreaction?, *Review of Financial Studies* 3, 175-205.

Moskowitz, Tobias J., and Mark Grinblatt, 1999, Do Industries Explain Momentum?, *Journal of Finance* 54, 1249-1290.

Menzly, Lior, and Oguzhan Ozbas, 2006, Cross-industry Momentum, Working Paper, University of Southern California.

Mitchell, Mark L., and J. Harold Mulherin, 1994, The Impact of Public Information on the Stock Market, *Journal of Finance* 49, 932-950.

Niederhoffer, Victor, 1971, The Analysis of World Events and Stock Prices, *Journal of Business* 44, 193-219.

Pritamani, Mahesh, and Vijay Singal, 2001, Return Predictability Following Large Price Changes and Information Releases, *Journal of Banking & Finance* 25, 631-656.

Roll, Richard, 1988, R^2, *Journal of Finance* 43, 541-566.

Rouwenhorst, K. Geert, 1998, International Momentum Strategies, *Journal of Finance* 53, 267-284.

Samuel, Arthur, 1959, Some Studies in Machine Learning Using the Game of Checkers, *IBM Journal of Research and Development* 3, 210-229.

Scholes, Myron, 1969, A Test of the Competitive Market Hypothesis: The Market for New Issues and Secondary Offerimgs, Unpublished Ph.D. thesis, Graduate School of Business, University of Chicago.

Shleifer, Andrei, and Lawrence H. Summers, 1990, The Noise Trader Approach to Finance, *Journal of Economic Perspectives*, 4(2): 19-33.

Tetlock, Paul C., 2007, Giving Content to Investor Sentiment: The Role of Media in the Stock Market, *Journal of Finance* 62, 1139-1168.

Tetlock, Paul, Maytal Saar-Tsechansky, and Sofus Macskassy, 2008, More Than Words: Quantifying Language to Measure Firms' Fundamentals, *Journal of Finance* 63, 1437-1467.

White, Halbert, 1980, Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity, *Econometrica* Vol. 48, No. 4, 817-838

Womack, Kent L., 1996, Do Brokerage Analysts' Recommendations Have Investment Value?, *Journal of Finance* 51, 137-167.

Wysocki, Peter, 1999, Cheap Talk on the Web: The Determinants of Postings on Stock Message Boards, Working Paper, University of Michigan Business School.

7.2 Digital sources

Factset

Kenneth French Data Library –

http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Nasdaq Stockholm – http://www.nasdaqomxnordic.com/

Riksbanken – http://www.riksbank.se/en/Statistics/

8 APPENDIX

Appendix I
Selection of online sources

Because of confidentiality reanson, we cannot disclose all online sources used.

https://www.flashback.org/nya-inlagg
https://www.avanza.se/placera/telegram.html
https://www.avanza.se/placera/pressmeddelanden.html
http://borsforum.svd.se/forum/6-aktier/
http://www.redeye.se/redeyes-analys
http://www.redeye.se/aktiebloggen
https://www.aktietorget.se/News.aspx
http://finwire.se/component/news
http://borssnack.di.se/
http://www.aktiespararna.se/forum/aktieanalys/
http://www.aktieexperterna.se/forum
http://www.bullstreet.se/debatt/forum/aktieforum/
http://www.sweshares.se/index.php
https://www.avanza.se/placera/aktier.html
https://www.aktietorget.se/finwire/News.aspx
http://www.axier.se/pressreleases
http://www.nyemissioner.se/nyheter/arkiv
http://introduce.se/
http://alternativa.se/Nyheter.aspx

Appendix II
Summary of sample stocks

203 Web Group	Dignitana	Karo Bio	Proact IT Group
Öresund	Dome Energy	Karolinska Development	Probi
AarhusKarlshamn	Doro	Kinnevik	Proffice
ABB	Drillcon	Klövern	Profilgruppen
Acando	Duni	Know IT	Ratos
Active Biotech	Duroc	Kopy Goldfields AB	RaySearch Laboratories
Addnode Group	East Capital Explorer	Kungsleden	Rejlerkoncernen
Addtech	Elanders	Lagercrantz Group	Rezidor Hotel Group
ADDvise Lab Solutions AB	Electra Gruppen	Lammhults Design Group	RNB Retail and Brands
Africa Oil Corp.	Electrolux	Latour	Rottneros
Agellis Group	Elekta	Lightlab Sweden	RusForest AB
Alfa Laval	Ellen	Lindab International	Sagax
Allenex	Elos	Loomis	Sandvik
AllTele	Endomines	Lucara Dimond Group	SAS
Anoto Group	Enea	Lundin Mining Corporation	SCA
Aqeri Holding	EnQuest PLC	Lundin Petroleum	ScandBook Holding AB
Arcam	Eolus Vind	Luxonen	Seamless Distribution
Arctic Gold AB	EOS Russia	Malmbergs Elektriska	Sectra
Arctic Paper	Episurf Medical AB	Meda	Securitas
Assa Abloy	Ericsson	MedCap	Semafo Inc.
AstraZeneca	Etrion	Medivir	Semcon
Atlas Copco	eWork Scandinavia	Mekonomen	Sensys Traffic
Atrium Ljungberg	EXINI Diagnostics AB	Melker Schörling	ShaMaran Petroleum Corp
Autoliv	Fabege	Micro Systemation	Shelton Petroleum
Avega Group	Fagerhult	Midsona	Sintercast
Axfood	Fast Partner	Midway	SJR
aXichem AB	Feelgood Svenska	Millicom International Cellular	Skåne-möllan
Axis	Fenix Outdoor	MQ Holding	SKF
B&B Tools	Fingerprint Cards	MSC Konsult	SkiStar
Balder	Firefly	MTG	Smarteq
Beijer Alma	FormPipe Software	multiQ International	Softronic
Beijer Electronics	G5 Entertainment	NAXS Nordic Access Buyout Fund	SSAB
Bergs Timber	Götenehus Group	NCC	Starbreeze
Betsson	Generic Sweden	Nederman Holding	Stille
Bilia	Genovis	Net Entertainment	Stora Enso
BillerudKorsnäs	Getinge	Net Insight	Studsvik
BioGaia	Geveko	NetJobs Group AB	Svedbergs
BioInvent International	Gränges	NeuroVive Pharmaceutical	Svolder
Biotage	Gunnebo	New Nordic Healthbrands	Sweco
Björn Borg	H&M	New Wave	Swedish Match
Black Earth Farming	Haldex	Nexam Chemical	Swedish Orphan Biovitrum
BlackPearl Resources	Hansa Medical	Nibe Industrier	Swedol
Boliden	Havsfrun Investment	Nischer AB	Systemair
Boule Diagnostics	Heba	Nobia	TagMaster
Bredband2	Heliospectra	Nolato	Tele2
Bringwell	Hemtex	Nordic Mines	TeliaSonera
BTS Group	Hexagon	Nordic Service Partners Holding	Tethys Oil
Bure Equity	Hexatronic Scandinavia AB	Note	Tieto
Byggmax Group	Hexpol	Novestra	Traction
C-RAD	Hifab	Novotek	TradeDoubler
Caperio Holding	HiQ International	Oasmia Pharmaceutical	Transcom WorldWide
Cassandra Oil	HMS Networks	Odd Molly	Transmode Holding
Castellum	Holmen	OEM International	Trelleborg
Catella	Hufvudstaden	Old Mutual	Trigon Agri
Catena	Husqvarna	Online Brands Nordic AB	Unibet Group
Cavotec	I.A.R Systems Group	Opcon	Uniflex
CellaVision	IFS	Opus Group	Unlimited Travel Group
ChronTech Pharma AB	Image Systems	Orexo	VBG Group
Clas Ohlson	Impact Coatings	Ortivus	Venue Retail Group
Cloetta	Industrivärden	Pallas Group AB	Victoria Park
Concentric	Indutrade	Paradox Entertainment	Vitec Software Group
Concordia Maritime	Insplanet	Peab	Vitrolife
Consilium	Intellecta	PetroGrand AB	Wallenstam
Corem Property Group	Intrum Justitia	Pfizer	WeSC
CTT Systems	Investor	Pilum	West International
CybAero	ITAB Shop Concept	PledPharma AB	Wihlborgs Fastigheter
Cybercom Group	Jays AB	Poolia	Wise Group
Dedicare	JLT Mobile Computers	Precio Systemutveckling	XANO Industri
DGC One	JM	Precise Biometrics	ZetaDisplay AB
Diös Fastigheter	KABE Husvagnar	Precomp Solutions	Zinzino
Diadrom Holding	Kancera AB	Prevas	
Diamyd Medical	KappAhl	Pricer	

Appendix III
Correlation matrix

This table shows the correlation matrix for the predictive regression (1–4). Weekday variables are excluded as they are binary.

	SSc	*Neg_Shock*	*Pos_Shock*	*Items*	$AbnRet_{t}$	$AbnRet_{t-1}$	$AbnRet_{t-2}$	*exMktret*	*SMB*	*HML*	*MOM*	*Log_turnover*
SSc	1.00											
Neg_Sh	-0.38	1.00										
Pos_Sh	0.66	0.21	1.00									
Items	-0.32	-0.46	-0.47	1.00								
$AbnRet_{t}$	-0.46	0.15	-0.18	0.07	1.00							
$AbnRet_{t-1}$	-0.09	0.08	-0.02	-0.08	0.06	1.00						
$AbnRet_{t-2}$	-0.07	0.07	-0.01	-0.04	0.04	0.02	1.00					
exMktret	0.05	-0.03	0.01	0.02	-0.07	-0.03	-0.10	1.00				
SMB	-0.02	0.02	0.01	0.02	0.08	-0.01	0.03	0.02	1.00			
HML	0.07	0.00	0.04	-0.01	-0.05	-0.14	0.02	0.06	0.41	1.00		
MOM	0.07	-0.01	0.04	-0.04	-0.09	0.02	0.01	-0.09	0.00	0.49	1.00	
Log_turnover	-0.01	-0.02	-0.04	-0.12	0.00	-0.08	-0.06	-0.02	-0.08	0.01	0.06	1.00

Appendix IV – VIF test

Controling for multicollinearity issues in predictive regressions (1–4)

This table shows the VIF statistics for four predictive regressions. The results show that the regressions are not materially influenced by multicollinearity.

Dependent variable	$exRet_{t+1}$		$exRet_{t+1}$		$AbnRet_{t+1}$		$AbnRet_{t+1}$
Sentiment proxy	*SSc*		*Neg_SSh* *Pos_SSh*		*SSc*		*Neg_SSh* *Pos_SSh*
HML	1.66	*Items*	1.69	*HML*	1.67	*Items*	1.68
Weekday_2	1.65	*HML*	1.66	*Weekday_4*	1.62	*HML*	1.67
Weekday_1	1.64	*Weekday_2*	1.65	*Weekday_2*	1.61	*Weekday_4*	1.62
Weekday_4	1.63	*Weekday_1*	1.64	*Weekday_3*	1.61	*Weekday_2*	1.61
Weekday_3	1.63	*Weekday_4*	1.63	*Weekday_5*	1.58	*Weekday_3*	1.61
UMD	1.41	*Weekday_3*	1.63	*UMD*	1.42	*Weekday_5*	1.58
SMB	1.26	*Pos_SSh*	1.43	*SMB*	1.26	*Pos_SSh*	1.43
Items	1.16	*UMD*	1.41	*Items*	1.16	*UMD*	1.42
SSc	1.13	*SMB*	1.26	*SSc*	1.13	*SMB*	1.26
Log_turnover	1.06	*Neg_SSh*	1.23	*Log_turnover*	1.06	*Neg_SSh*	1.23
Ex_market_return	1.03	*Log_turnover*	1.05	*Ex_market_return*	1.04	*Log_turnover*	1.05
$AbnRet_{t-1}$	1.03	*Ex_market_return*	1.03	$AbnRet_{t-1}$	1.03	*Ex_market_return*	1.04
$AbnRet_{t}$	1.03	$AbnRet_{t-1}$	1.03	$AbnRet_{t}$	1.03	$AbnRet_{t-1}$	1.03
$AbnRet_{t-2}$	1.02	$AbnRet_{t}$	1.03	$AbnRet_{t-2}$	1.02	$AbnRet_{t}$	1.03
		$AbnRet_{t-2}$	1.02			$AbnRet_{t-2}$	1.02
Mean VIF	1.31		1.36		1.30		1.35

Appendix V
Correlation matrix

This table shows correlation matrix for variables used in trading strategy regression.

	R_{Mkt}	*SMB*	*HML*	*MOM*	*Intercept*
R_{Mkt}	1.00				
SMB	-0.28	1.00			
HML	0.08	0.01	1.00		
MOM	0.13	-0.15	0.36	1.00	
Intercept	-0.04	-0.09	0.06	-0.04	1.00

Appendix VI
VIF test – controling for multicollinearity issues in regression 5

This table shows the VIF statistics for trading strategy regression. The results show that the regression is not materially influenced by multicollinearity.

Dependent variable	*R_Port*
HML	1.65
UMD	1.4
SMB	1.22
R_{Mkt}	1.02
Mean VIF	1.32