# Numerisization of Investor Sentiment in News and Application to Stock Reactions

Master's Thesis submitted

to

**Prof. Dr. Wolfgang Karl Härdle**

Humboldt-Universität zu Berlin

School of Business and Economics

Institute for Statistics and Econometrics

Ladislaus von Bortkiewicz Chair of Statistics

by

**Elisabeth Bommes**

552875



in partial fulfillment of the requirements

for the degree of

**Master of Science**

Berlin, March 1, 2015

## Abstract

In an era where information is publicly available on the internet and computers are able to handle large amounts of data, it is only logical to utilize news and other text sources to improve our understanding of stock reactions. Namely, these stock reactions are volatility, trading volume and return. This thesis proposes a guide line on how to extract and process stock related information from different sources such as news and investment articles. Furthermore, sentiment is projected by using a finance specific lexicon that classifies words in either positive or negative. Hence, the sentiment is numerisized and it is possible to utilize it in mathematical models such as panel regression. The derived sentiment is then used to answer the following research questions:

(i) Does the nature of the derived sentiment measure play an important role?
(ii) Is there an asymmetric response given the sentiment values?
(iii) Is there evidence for the validity of the uncertain information hypothesis?

We find that the choice of sentiment measure is indeed important for the subsequent analysis. Also, empirical evidence points in the direction that there is in fact an asymmetric response. Furthermore, the incremental information in the distilled news flow is in line with the information hypothesis.

*Keywords: Sentiment Measures, Investor Sentiment, News Analysis, Volatility, Trading Volume, Returns*
*JEL Classifications:* C80, C81, G02, G14

# Contents

**List of Figures** **5**

**List of Tables** **6**

**1 Introduction** **1**

1.1 Literature . . . . . 2

1.2 Outline . . . . . 4

**2 Web Scraping** **6**

2.1 Basics of Web Scraping . . . . . 6

2.2 Article Links . . . . . 7

2.3 Article . . . . . 8

2.4 Legality of Web Scraping . . . . . 9

**3 Sentiment Extraction** **11**

3.1 Sentiment Lexica . . . . . 11

3.2 Processing Steps . . . . . 15

3.2.1 Text Cleaning . . . . . 15

3.2.2 Tokenization . . . . . 17

3.2.3 Chunking . . . . . 18

3.2.4 Slicing . . . . . 19

3.3 Sentiment Tagging . . . . . 20

**4 Data** **23**

4.1 Financial Articles . . . . . 23

4.2 Stock Data . . . . . 26

**5 Empirical Results** **29**

5.1 Contemporaneous Regression . . . . . 29

5.2 Time-Lagged Panel Regression . . . . . 33

**6 Conclusion** **35**

**Appendices** **38**

**A Article Example** **38**

**B Additional Tables** **40**

**References** **42**

# List of Figures

1 Data Gathering and Processing Steps . . . . . 5
2 Monthly Correlation between Positive Sentiment . . . . . 14
3 Monthly Correlation between Negative Sentiment . . . . . 15
4 S&P 500 Articles per Day . . . . . 23
5 Example of an Article . . . . . 39

# List of Tables

1 Most frequent Words . . . . . 12
2 Pairwise Comparison of Lexica . . . . . 13
3 Derived Sentiment Variables . . . . . 21
4 Frequencies of Stock Symbols . . . . . 24
5 Summary Statistics . . . . . 27
6 Contemporaneous Regression Results on Word Level . . . . . 30
7 Contemporaneous Regression Results on Word Level: Model 4 and 5 . . . . 31
8 Time-Lagged Panel Regression Results on Word Level . . . . . 36
9 Time-Lagged Regression Results on Word Level: Model 4 and 5 . . . . . 37
10 Contemporaneous Regression Results on Sentence Level . . . . . 40
11 Time-Lagged Panel Regression Results on Sentence Level . . . . . 41

# 1 Introduction

Since the groundbreaking work of Bachelier (2006) in his doctoral thesis which is dating back to 1900, it is common knowledge that nobody is able to predict tomorrow's stock price by looking at the history of prices. But it is also widely accepted that information flow is a huge contributor in the price adaption process of financial assets. This is especially valid for volatility and trading volume as news play an important role in the theoretical model of realized volatility as stated in Andersen et al. (2003).

However, it is not as clear whether new information also helps to explain or even predict the stock price in the next period due to the efficient market hypothesis (EMH). The strong-form of the EMH by Fama (1970) suggests that all information is directly reflected by the share prices and no market participant can earn excess returns. Its weak-form claims that stock prices incorporate all prior public information and investors cannot gain excess returns by using a trading strategy that relies on historical information. Hence, if the EMH holds, then today's stock returns should not be correlated to yesterday's sentiment in investment specific articles.

There is empirical evidence that the EMH does not hold in reality and a survey regarding this topic is provided by Malkiel (2003). Thus, the uncertain information hypothesis (UIH) was developed by Brown et al. (1988) based on the overreaction hypothesis by Bondt and Thaler (1985). In contrast to the EMH, the market participants set new prices before the full range of the news content is resolved. In case of both favorable and unfavorable news, the investors set stock prices significantly below their conditional expected values and thus, react risk-averse.

Bondt and Thaler (1985) provide evidence for this overreaction hypothesis as prior losing stocks historically earn about 25% more than the former winners. They do not rely on sentiment in news texts for their analysis but define a "news event" as extreme movement in the stock prices. Zarowin (1990) argue that the overreaction phenomen is a result of the size effect due to the reduced market capitalization of the loser stocks. The size effect is based on the work of Banz (1981) who report that historically, firms with smaller market capitalization consistently outperform stocks with larger market capitalization. Numerous other researchers such as Brown et al. (1988) and Yu et al. (2010) provide further evidence for the UIH. However, this prior work depends on news measures that are directly derived from stock reactions, namely returns, while news sentiment specific work does not re-examine the UIH.

In an era where information is publicly available on the internet and computers are able to handle large amounts of data, it is only logical to utilize news and other text sources to improve our understanding of stock reactions. Namely, these stock reactions are volatility, trading volume and return. This thesis proposes a guide line on how to extract and process stock related information from different sources such as news and investment articles. Furthermore, sentiment is projected by using a finance specific lexicon by Loughran and McDonald (2011) that classifies words in either positive or negative. Hence, the sentiment is numerisized and it is possible to utilize it in mathematical models such as panel regression. The derived sentiment is then used to answer the following research questions:

(i) Does the nature of the derived sentiment measure play an important role?

(ii) Is there an asymmetric response given the sentiment values?

(iii) Is there evidence for the validity of the UIH?

Question (i) aims at the projection of sentiment on numerical values. For example, good and bad sentiment can either be condensed in one measure or be treated seperately. However, the results should still point in the same direction if the value of sentiment is robust. Question (ii) addresses the well known fact that e.g. volatility increases more if past returns are negative. Hence, it should also react stronger to bad news than good news. Due to previous work by Zhang et al. (2015), it can be expected that there is an asymmetric response and the market should react stronger to negative news. Question (iii) aims to fill a gap in previous research. While many researchers have presented evidence for the validity of the UIH, to the author's knowledge none of them have used textual news but they have derived "news events" directly from the stock prices.

## 1.1 Literature

Traditional news sources such as Wall Street Journal articles are widely used in sentiment and news analysis. As previously mentioned, there are numerous studies that look at the relationship between news flow and stock reactions but many of them do not regard a textual analysis to isolate sentiment. Tetlock (2007) concludes that negative sentiment in a Wall Street Journal column has explanatory power for downward movement of the Dow Jones. Wisniewski and Lambe (2013) collect news from the Lexis-Nexis database and filter

for phrases that were dominantly used during the credit crisis: "Credit Crunch", "Financial Crisis" and "Bank Failures". Their findings suggest that news might influence future market movements while there is only weak evidence that journalists repeat prior news. Groß-Klußmann and Hautsch (2011) analyze market reactions to the intra-day stock specific data from the "Reuters NewsScope Sentiment" engine with already derived sentiment values. Their findings support the hypothesis that news influences volatility and trading volume but are limited to a small number of assets due to their high frequency context.

Early work in sentiment analysis in social media with application to stock markets often focuses on message boards such as *Yahoo! Finance* and *Raging Bull* as text source. Advantages of stock message boards as data source are that new messages are posted frequently and also, the discussed stocks can easily be identified. Antweiler and Frank (2004) analyze text contributions from the two mentioned stock message boards and find that the amount and bullishness of messages have predictive value for trading volume and volatility. Das and Chen (2007) conclude that there is a positive correlation between aggregated message board sentiment and the next day's stock index return. However, they do not find evidence on individual firm level. On message boards, the self-disclosed sentiment to hold a stock position is not bias free, as indicated in Zhang and Swanson (2010).

Sabherwal et al. (2011) investigate stock market manipulations due to "pump and dump" strategies in message boards of small-cap firms. They find a pattern that suggests the possibility to manipulate small firm's stock prices via online discussion. Park et al. (2013) assess the impact of stock message boards on real life investors and suggest that there is a confirmation bias. Hence, the traders prefer messages that support their prior beliefs. A disadvantage of message board data is that the history of many collected samples is quite limited. This might be due to the fixed number of messages that is shown on *Yahoo! Finance* instead of the whole history. For example, Antweiler and Frank (2004) and Sabherwal et al. (2011) only use one year of message posting at maximum.

A larger sample that spans six years is analyzed by Bettman et al. (2011). They use Naive Bayes classification to filter for potential takeover rumors and subsequently find that abnormal returns and trading volumes follow the posting of these rumors. Another study that spans six years of message posting is provided by Kim and Kim (2014). They find evidence that prior stock price performance influences future message board posting instead of the other way around. Li et al. (2014) extract news and message board postings

from Chinese web pages and their "electronic-media-aware quantitative trader" algorithm outperforms the Chinese CSI 100.

In recent years, the micro-blogging platform *Twitter* gained popularity in sentiment research. Twitter messages (tweets) are limited to 140 characters as they are usually sent by a mobile device. Thus, grammar and multiple mentioned stocks are usually no hindrance in the analysis. Bollen et al. (2011) classify tweets in six different mood states and suggest that public mood helps to predict changes in Dow Jones values on a daily level. Zhang et al. (2012) refine the filtering process by isolating keywords indicating a financial context. They also consider different markets such as commodities and currencies. Si et al. (2013) extend the filtering to obtain tweets on firm level and conclude that the extracted topic based Twitter sentiment improves day-to-day stock forecast accuracy. Sprenger et al. (2014) also use tweets on stock level and conduct that the number of followers and retweets may be used to successfully evaluate the quality of investment advice. Nann et al. (2013) aggregate data from stock message boards, *Twitter* and traditional news sources and outperform the S&P 500 with their sentiment trading model.

Even more recent, researchers use articles and comments from social media related investment communities such as *Seeking Alpha*. Chen et al. (2014) use articles and corresponding comments on *Seeking Alpha*, a social media platform for investment research, and show predictive value of negative sentiment for stock returns and earnings surprises. According to Wang et al. (2014), the correlation of Seeking Alpha sentiment and returns is higher than between returns and sentiment in Stocktwits, messages from a micro-blogging platform specialized in finance. Zhang et al. (2015) obtain articles from *NASDAQ Community*, a platform that gathers articles from various news and social investing platforms. They compare different sentiment lexica and find an incremental impact of the derived sentiment on returns, trading volume and volatility.

## 1.2 Outline

The further work is structured as follows. Section 2 and 3 introduce the steps to gather and process the data. These steps are summarized in Figure 1. Firstly, news articles are automatically downloaded by a computer program, later introduced as web scraper in Section 2. As the raw text material cannot be used in statistical models, further processing steps to extract sentiment are necessary and introduced in Section 3.

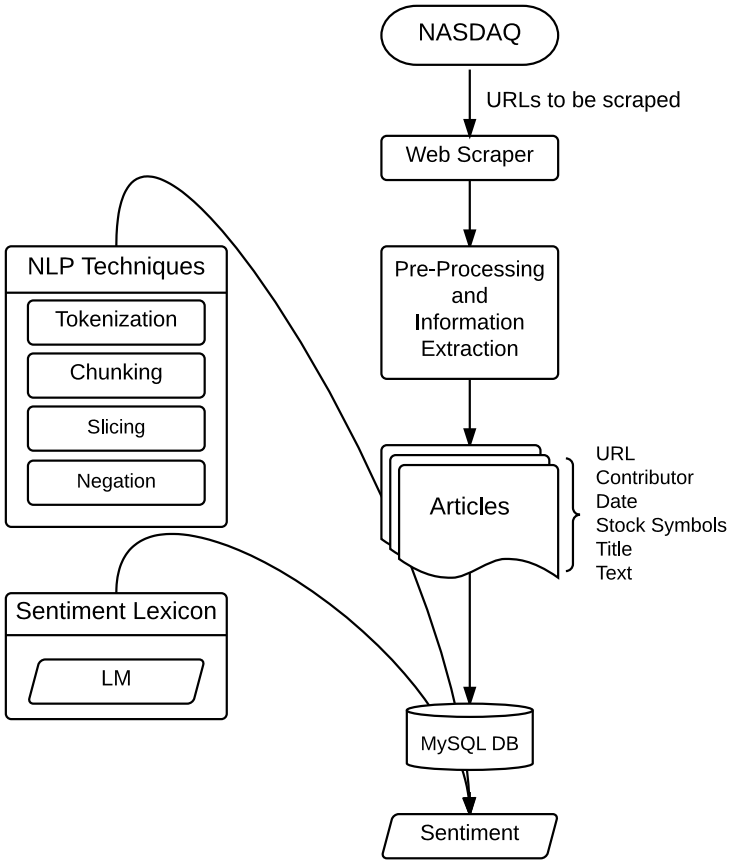


**Figure 1:** Flowchart of Data Gathering and Processing Steps

The used sentiment lexicon is discussed in Section 3.1. Section 3.2 contains information on how the text is cleaned and treated. Firstly, the text is cleaned in 3.2.1. Then, in 3.2 specific information (meta data) like the mentioned stock symbol is stripped from the article and stored seperately. Tokenization is used in 3.2.2 to break the text in usable units. Afterwards, company names are aggregated in units by using a chunking technique in 3.2.3 and furthermore, the text is sliced to account for multiple companies in a single article in 3.2.4. Then, the actual classification of sentiment is done in 3.3.

Section 4 summarizes the used data sets, both financial articles in 4.1 and financial variables in 4.2. The empirical analysis by with panel regression models follows in Section 5 while Section 6 concludes.

# 2 Web Scraping

Massive amounts of data are available in the web. Too much, to be filtered and analyzed manually. As the information is typically not in a machine-readable format, automatic programs are built to parse the documents and extracting specific data points. These programs are commonly called web scraper. In this section, basics of web scraping are explained, the implemented web scraper is discussed and some remarks on the legality of web scraping are given.

## 2.1 Basics of Web Scraping

While this section just covers the absolute basics of web programming to understand how the automatic extraction of information works, there are various introductory books on web page programming. For instance, Duckett (2011) and Lane (2012) are good books to begin with. There are generally three standard web programming languages that are widely used to program web pages: HyperText Markup Language (HTML), Cascading Style Sheets (CSS) and Javascript (JS).

```
<!DOCTYPE html>
<html>
<head>
<title>The Header</title>
<style type="text/css">
...
</style>
...
</head>
<body>
<script language="javascript">...</script>
<h1>This is a large heading</h1>
<p>Text...</p>
<h3>This is a smaller heading</h3>
<div>
<p>More text...</p>
</div>
</body>
</html>
```

**Listing 1:** Example HTML Code with embedded CSS and JS elements

HTML is used to structure a web document while CSS and JS are used to control the styling and creating dynamic content, respectively and normally embedded in the HTML code. An example of HTML code can be seen in listing 1.

As everything in between the HTML tags `<head>` and `<\head>` is just meta data that is not explicitly seen by the reader , the focus lies on the so called body of the HTML document that is embedded by `<body>` and `<\body>`. Typically, the tag `<p>` denotes standard text while tags like `<h1>` to `<h6>` denote different headings. To access specific information on a web page, it is essential to parse the HTML code. A parser, in this sense, just analyzes the input code and imposes a structure that allows to travel to every specific node or path in the code. After the parsing, it is possible to access specific nodes in the code; e.g. if you want to have access to the large heading `<h1>` you can just go to the path `<html><body><h1>` and you get "This is a large heading" as output.

The HTML file can now be filtered for information between specific tags by searching for nodes between these tags. As long as the article is surrounded by specific and identifiable tags, we are able to gather data in an automatic way.

### 2.2 Article Links

NASDAQ.com has a subpage with links to the latest articles and the goal of this section is to scrape the link of each article that has been published. Firstly, we notice that it is possible to iterate over the subsequent pages of `latest-articles.aspx` by attaching the term `?page=i` to the link where `i` is the iterator. Hence, we can obtain the link to every article published on NASDAQ Community as long as we are able to scrape the links on a single page.

In HTML, a link is introduced by the tag `<a href='PATH' ...>LINK TEXT </a>` where `PATH` is the part we are interested in as it is the actual link to the document. Hence, we want to access the attribute `href` of the tag `a`. This, of course, would return every link on a specific internet page and thus, also the page navigation. As we are only interested in links to articles, we need an additional identifier for links to articles.

```
<!DOCTYPE html>
<html lang="en-us" class="inner no-js" xmlns:og="http://ogp.me/ns#" xmlns:fb="
    https://www.facebook.com/2008/fbml">
<head>...</head>
<body>
        ...
    <h1>Latest Articles</h1>
```

```
    <div style="clear: both;"></div>
    <div class="article"><a href="http://www.nasdaq.com/author/fool"><img class
        ="la_avatar" src="/common/images/TMFJester40x40.png" alt="Motley Fool"
        /></a><p><strong><a href="http://www.nasdaq.com/article/walgreen-
        company-delivers-a-prescription-for-growth-cm426543">Walgreen Company
        Delivers a Prescription for Growth</a></strong></p><p class="
        la_articleinfo">12/23/2014, 10:39 pm from <a href="http://www.nasdaq.
        com/author/fool">Motley Fool</a> in <a href="http://www.nasdaq.com/
        investing/">Investing</a>, <a href="http://www.nasdaq.com/investing/
        stocks.stm">Stocks</a></p><div style="clear:both;"></div><p>    Source
         Walgreen.  Walgreen Company is the nation's biggest pharmacy store
        operator, with more than 8,200 stores, but it's also becoming a major
        global drug retailer thanks to its merger with Alliance Boots, which...
         <a href="http://www.nasdaq.com/article/walgreen-company-delivers-a-
        prescription-for-growth-cm426543">Read &gt;&gt;</a></p></div><div class
        ="article"> ... <div> ...
    ...
</body>
</html>
```

**Listing 2:** Part of Source Code of "Latest Articles"; complete Source Code has more than 1000 Lines

By looking at part of the source code in Listing 2, we can see that the links to an article are always in a block that is introduced by the tag `<div class="article">`. Furthermore, article links have the pattern `/article/` in contrast to other links, e.g. the link to the page about the contributor which has the pattern `/author/`. All links to articles are scraped and stored in a MySQL database. Furthermore, to avoid duplicates a restriction is used such that each link in the database must be unique. The corresponding program to scrape all links can be found in TXTScrapeLinks.R .

### 2.3 Article

After the scraping of all links to articles, it is quite easy to obtain the content of each article. However, due to navigation and pictures, each article needs roughly 4 megabyte (MB) disc space and scraping each page completely would not be efficient. Furthermore, as the page structure is repeated throughout the articles, this would not lead to additional information. However, the print version of each article can be downloaded at a fixed link (`stockmarketnewsstoryprint.aspx?storyid=`) by using the article id which

is part of the already scraped links. This reduces the size of the HTML document to less than 1 MB. One example of an article's print version can be seen in the Appendix (A) in Figure 5. Important information about the article, namely metadata, is directly given at the top of the article: title, contributor date, referenced stocks and the actual article text. Due to the HTML markup, these parts can be easily identified. The corresponding identifiers are:

- `<h1 itemprop="headline">` for the title,
- `<div id="by-line">` for the contributor,
- `<div itemprop="dateCreated" id="articledateposted">` for the date,
- `<div id="referenced-stocks">` for the referenced stock symbols and
- `<div id="articleText">` for the actual article text.

The corresponding program to scrape all articles and extract meta data can be found in TXTScrapeArticles.R . In the next step, the article text is cleaned from any HTML, JS and CSS markups and the text is stored, together with the meta data, in the MYSQL database.

### 2.4 Legality of Web Scraping

The literature regarding the legality of web scraping is quite sparse. Both Jennings and Yates (2009) and Truyens and Van Eecke (2014) provide an overview about legal aspects of web scraping. As Truyens and Van Eecke (2014) states the current state of the art from the perspective of the textminer, some of their conclusions are summarized below. Generally, content of web pages is protected by copyright law both in the United States (US) and the European Union (EU) such that the content may not be republished. Non-commercial academic research is an exception of this rule such that web scraping is principally legal in this context. However, the Terms of Service (ToS) of a web page are still binding. This is often problematic as web page owners might prohibit the use of web scrapers or other automatic programs to gather information from their page. As an example, *Seeking Alpha* prohibits the use of any automated program to "scrape" or "harvest" their page. *Yahoo! Finance*, another popular page in sentiment analytics does not include such a term in their ToS, however, the limited message history might be an obstacle. As of December 2014, only the last 10,000 messages are shown in each stock specific message board. This more or less corresponds to a two-month-period for stocks that people frequently talk about like *Apple*. Contrary to the previously mentioned pages, NASDAQ offers a platform for

financial articles by selected contributors including social media websites such as *Seeking Alpha* and *Motley Fool* as well as investment research firms such as *Zacks*. Their ToS do not prevent the utilization of a web scraper and the article history is not limited.

# 3 Sentiment Extraction

Text is an unstructured data source. This means that the text may contain intelligence about investor sentiment but this information has to be extracted as it is not provided in grouped or numerical form. Thus, it is necessary to use processing steps to convert the raw text in a format that allows the identification of sentiment signals. The identification and classification of sentiment in text can then be done by using a specialized lexicon. After the identification of sentiment in the articles, a function is needed to measure the sentiment and thus, map it to a value that can be used in the further analysis. Section 3.1 presents widely used sentiment lexica. Section 3.2 provides an overview of the used processing steps.

## 3.1 Sentiment Lexica

There are several sentiment lexica freely available which can be used to assign positive or negative sentiment to text. Both Wang et al. (2014) and Chen (2014) use the financial sentiment lexicon (LM) by Loughran and McDonald (2011). The LM lexicon is derived from annual 10-K filings of U.S. companies. Contrary to more general dictionaries such as the Harvard Psychosociological Dictionary, the LM lexicon does not contain in a financial sense ambiguous words such as *tax* or *board*. To give an example, interpreting the word *board* as negative would lead to a misclassification of sentiment every time an article mentions the phrase *board of directors*. The LM lexicon contains 354 positive, 2,329 negative, 297 uncertainty, 886 litigious, 19 strong modal and 26 weak modal words. However, words in the uncertainty list might also appear in the negative word list and hence, publications such as Zhang et al. (2015) and Chen (2014) only consider positive and negative words. Zhang et al. (2015) compare the LM lexicon extensively to the opinion lexicon (BL) by Hu and Liu (2004) and the Multi-Perspective Question Answering subjectivity lexicon (MPQA) by Wilson et al. (2005). The BL lexicon lists 6,789 words (2,006 positive and 4,783 negative) while the MPQA lexicon consists of 8,222 entries. Additionally to the processing steps in Section 3.2, the application of the MPQA lexicon requires part-of-speech tagging and stemming of the words in each article. All of these lists only regard single word terms and multi word terms like bigrams or even trigrams are not covered.

As Zhang et al. (2015) use the same data source, some of their findings are presented here. By taking a look at the unique words of each lexicon that appear at least three

| BL | | LM | | MPQA | |
|---|---|---|---|---|---|
| Positive (470) | Negative (918) | Positive (267) | Negative (916) | Positive (512) | Negative (181) |
| Available | Debt | Opportunities | Declined | Just | Low |
| (5,836) | (12,540) | (4,720) | (9,809) | (17,769) | (12,739) |
| Led | Fell | Strength | Dropped | Help | Division |
| (5,774) | (9,274) | (4,393) | (4,894) | (17,334) | (5,594) |
| Lead | Fool | Profitability | Late | Profit | Least |
| (4,711) | (5,473) | (4,174) | (4,565) | (15,253) | (5,568) |
| Recovery | Issues | Highest | Claims | Even | Stake |
| (4,357) | (3,945) | (3,409) | (3,785) | (13,780) | (4,445) |
| Work | Risks | Greater | Closing | Deal | Slightly |
| (3,808) | (2,850) | (3,321) | (3,604) | (13,032) | (3,628) |
| Helped | Issue | Surpassed | Closed | Interest | Close |
| (3,631) | (2,821) | (2,464) | (3,378) | (12,237) | (3,105) |
| Enough | Falling | Enable | Challenges | Above | Trial |
| (3,380) | (2,768) | (2,199) | (2,574) | (12,203) | (2,544) |
| Pros | Aggressive | Strength | Force | Accord | Decrease |
| (2,841) | (1,796) | (2,157) | (2,157) | (11,760) | (2,205) |
| Integrated | Hedge | Alliance | Unemployment | Natural | Disease |
| (2,652) | (1,640) | (1,842) | (2,062) | (10,135) | (2,001) |
| Savings | Proprietary | Boosted | Question | Potential | Little |
| (2,517) | (1,560) | (1,831) | (1,891) | (9,905) | (1,775) |

**Table 1:** Lists of ten most frequent positive and negative Words that are unique in each Lexicon, Source: Zhang et al. (2015)

times in the NASDAQ articles, it can be easily seen that there are 1,388 unique words in the BL lexicon (470 positive and 918 negative), 1,183 unique words in the LM lexicon (267 positive and 916 negative) and 693 unique words in the MPQA lexicon (512 positive and 181 negative). Table 1 shows the ten most frequent positive and negative words that are unique to the BL, LM and MPQA lexicon. The words in the BL and MPQA lexicon seem to be more general than the words in the LM lexicon. Other words, like *Profit* in the MPQA lexicon have a connection to finance but appear quite often in the articles and might be ambiguous without regarding the specifying adjective. Thus, the word *Profit* itself does not relate to profitability without checking whether the profits are high or low. Another problem could arise by identifying the word *Fool* als negative as does the BL lexicon. One of the contributors is named *Motley Fool* which is why this term quite often appears in the provided articles as there is often a link to the web page on the bottom of

articles. However, similar situations with problematic ambiguity can also be constructed for the LM lexicon as the word *Closing* might just reference the closing price instead of the closing of a company. Here must be stated that the content of the yearly 10-K filings does not totally correspond to the content of daily news articles. Thus, the choice of wording and phrases might be different. Zhang et al. (2015) also look at the pairwise similarities

| BL and LM | | BL and MPQA | | LM and MPQA | |
|---|---|---|---|---|---|
| Positive (131) | Negative (322) | Positive (971) | Negative (1164) | Positive (32) | Negative (30) |
| Gains | Losses | Free | Gross | Despite | Against |
| (7,604) | (5,938) | (133,395) | (8,228) | (7,413) | (8,877) |
| Gained | Missed | Well | Risk | Able | Cut |
| (7,493) | (3,165) | (30,270) | (7,471) | (5,246) | (3,401) |
| Improved | Declining | Like | Limited | Opportunity | Challenge |
| (7,407) | (3,053) | (24,617) | (5,884) | (4,398) | (1,042) |
| Improve | Failed | Top | Motley | Profitable | Serious |
| (5,726) | (2,421) | (14,899) | (5,165) | (3,580) | (1,022) |
| Restructuring | Concerned | Guidance | Crude | Efficiency | Contrary |
| (3,210) | (1,991) | (11,715) | (5,109) | (2,615) | (401) |
| Gaining | Declines | Significant | Cloud | Popularity | Severely |
| (3,150) | (1,654) | (10,576) | (4,906) | (1,588) | (348) |
| Enhance | Suffered | Worth | Fall | Exclusive | Despite |
| (2,753) | (1,435) | (10,503) | (4,732) | (1,225) | (342) |
| Outperform | Weaker | Gold | Mar | Tremendous | Argument |
| (2,518) | (1,288) | (9,303) | (3,190) | (611) | (324) |
| Stronger | Critical | Support | Hard | Dream | Seriously |
| (1,657) | (1,131) | (9,120) | (2,957) | (581) | (240) |
| Win | Drag | Recommendation | Cancer | Satisfaction | Staggering |
| (1,491) | (1,095) | (8,993) | (2,521) | (410) | (209) |

**Table 2:** Pairwise Comparison of Lexica: 10 most frequent positive and negative Words. Source: Zhang et al. (2015)

of the lexica. BL and LM share 131 positive and 322 negative words that are not part of the MPQA lexicon. With 2,135 shared words (971 positive and 1,164 negative) the intersection of BL and MPQA (without LM words) contains more positive words than the whole LM lexicon. LM and MPQA only share 32 positive and 30 negative terms that do not also belong to the BL lexicon. Table 2 shows the most frequent terms that are part of two dictionaries, but not all three lexica. It can be directly seen that the each of the ten most frequent positive words that are shared by BL and MPQA appear far more often than

the most frequent positive word that is part of LM and BL or LM and MPQA. Also, it is questionable whether the word *Free* can be a meaningful indicator of positive sentiment as it appears 133,395 times which is more often than the following nine words together. Also, words like *Gold* and *Crude* (oil) probably cannot be seen as positive or negative as they just refer to investment products. The word *Cancer* is probably sector specific as it should appear more often in the Health Care area. But again, misclassification is also possible for the LM lexicon as an argument does not have to refer to a dispute but can also refer to a reason to buy stocks of a specific company.

Nonetheless, Zhang et al. (2015) find that the sentiment measures of the three lexica are highly positive correlated over time. Figure 2 and 3 show the monthly positive and negative correlation, respectively. Generally, the correlation between the lexica is higher for negative sentiment. The results for the different lexica are consistent with the results of Tables 1 and 2 as the correlation of BL and MPQA is higher than the correlation of LM with these lexica. Also, the correlation seems to be more stable over time in recent years than in the time before 2012. The further focus on this thesis lies on the LM lexicon as

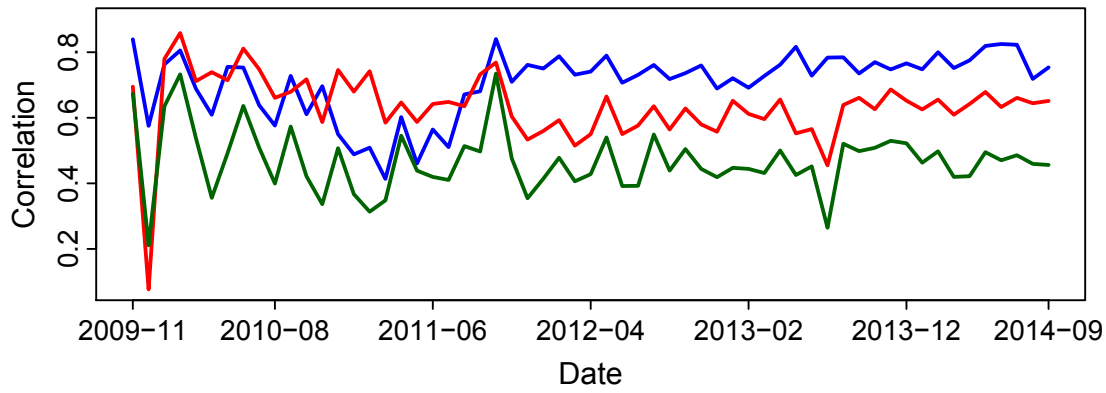


**Figure 2:** Monthly Correlation between Positive Sentiment: BL and LM, BL and MPQA, LM and MPQA. Source: Zhang et al. (2015)

an extensive comparison of different lexica has already been done by Zhang et al. (2015). An alternative to the fixed word lists in sentiment lexica are classification techniques like Naive Bayes or a Support Vector Machine based approach. This is often done for shorter and straight forward texts like stock board or Twitter messages. The application of such algorithms can be seen in Antweiler and Frank (2004) and Wang et al. (2014). However, training data is needed and this requires manual labeling of large text portions. Fixed

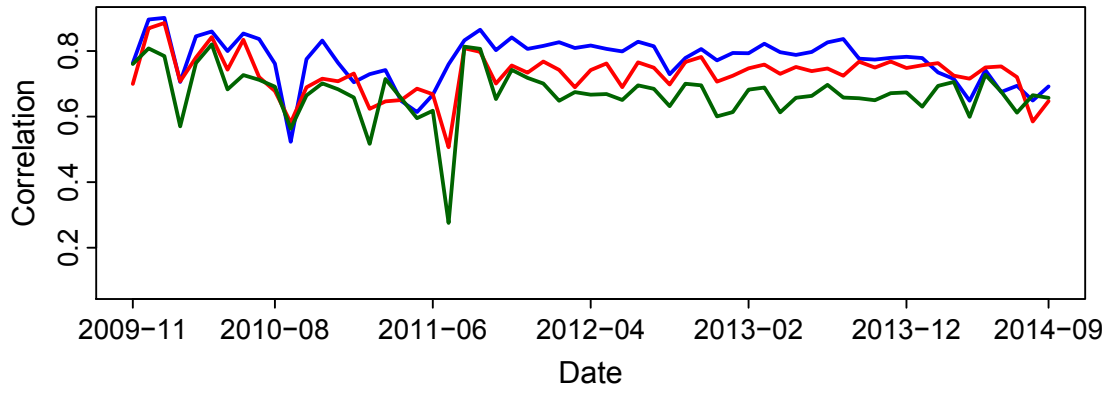


**Figure 3:** Monthly Correlation between Negative Sentiment: BL and LM, BL and MPQA, LM and MPQA. Source: Zhang et al. (2015)

word lists also have the advantage that overfitting is not possible and the results are comparable across data sets.

### 3.2 Processing Steps

Each article as a whole can be seen as semi structured due to easily exctractable information such as the date, contributor and referenced stock symbols. This is not the case for the article text which is just a sequence of characters. Of course, it would be possible to identify how many parts of this sequence are identical to terms in the LM lexicon but this way the word *unbroken* would be identified as *broken* which is a negative word in the LM lexicon. To avoid errors like this, we have to look at the implicit structure of text which consists, among other things, out of words, numbers and punctuation. Natural Language Processing (NLP) techniques are used to transform this implicit structure into an explicit structure. The NLP is partly done by using the Natural Language Processing Toolkit (NLTK) by Bird et al. (2009) in the programming language Python.

#### 3.2.1 Text Cleaning

The original text source is encoded in Unicode and more specifically in the Universal Character Set with Transformation Format 8 bit (UTF-8). This basically means that the used character set allows for language specific characters like *ä* in German or *ø* in Danish. As the articles are in English, Unicode characters are not as relevant as stated in Bird et al. (2009). Trial runs for this study have shown that UTF-8 characters may lead to

errors while executing the source code.

```
def clean_up(txt):
        txt = str(txt.encode('ascii', 'ignore'))
        txt = txt.replace("\\n"," ")
        txt = txt.replace("\n"," ")
        txt = txt.replace('b"',"")
        return txt
```

**Listing 3:** Source Code to set change Characterset from UTF-8 to ASCII

TXTProcessing.py

Hence, the text with UTF-8 encoding is converted to American Standard Code for Information Interchange (ASCII) while removing characters that do not exist in the ASCII set. The relevant source code can be seen in Listing 3. Additionally indicators for new lines, namely `\n` and `\\n` are replaced by a white space.

While capitalization, numbers and punctuation characters deliver more information for humans, a machine does not know that a word with and without suceeding period may have the same meaning. Thus, a short Python function is presented in Listing 4 to standardize the text.

```
def remove_punctuation(txt):
        punct_numb = '"#$%&\()*+,-./:;<=>?@[\\]^_`{|}~€' + '0123456789'
        txt_without_punct = ''
        for letter in txt:
                if letter not in punct_numb:
                        txt_without_punct += letter.lower()
                else:
                        txt_without_punct += ' '
        return txt_without_punct
```

**Listing 4:** Set characters to lower case, remove punctuation and numbers

TXTProcessing.py

The variable `txt` specifies the input string while `punct_numb` contains all characters that should be removed from the text. Then, the function loops over each letter in `txt` and checks whether it is part of `punct_numb`. If this is not the case, the letter is set to lower case and is attached to the variable `txt_without_punct` which is then returned after the loop.

### 3.2.2 Tokenization

The process of breaking text down into these basic units is called word tokenization. While it is due to space delimiters fairly easy to identify word tokens in English, it is impossible to automatically analyze the text without tokenization (Webster and Kit, 1992). However, the short example sentence `("Apple's new iphone is now in the stores.")` can be seen as vector with one entry, `(.)` specifying the vector. The sentence illustrates that it is not sufficient to use space delimiters to identify word tokens as this would lead to `("Apple's", "new", "iphone", "is", "now", "in", "the", "stores.")`. Hence, it does not break `"Apple's"` into `"Apple", "'s"` and would not allow to automatically detect that the sentence is about the company Apple. Also, punctuation is not treated as seperate token. By contrast, NLTK incorporates Penn Treebank tokenizer by MacIntyre (1995) and is able to correctly split the Anglo-Saxon genitive of nouns.

The Penn Treebank tokenizer requires that the sentence boundaries are already detected which is referred to as sentence tokenization. Similar to the word tokenization, the naive approach would be to split the text into sentences by identifying punctuation such as `"."`, `"?"` and `"!"` to use them as boundaries. But it must be noted that abbreviations such as `"Mr."`, initials and some numbers are also followed by a period. Also ellipses to mark the intentional omission of a word (marked by `"..."`) must be considered as exception. Furthermore, if an abbreviation is the last word of a sentence, there is only one period which then marks the abbreviation and the end of the sentence at the same time. Hence, sentence boundary detection must disambiguate these cases.
Kiss and Strunk (2006) construct the Punkt tokenizer, an unsupervised approach to detect these sentence boundaries in a text.
The disambiguation in their algorithm takes place in two stages:

(i) Type-based stage: Initial phase where abbreviations and ellipses are detected.

(ii) Token-based stage: The annotation of the initial phase are corrected by detecting abbreviations and ellipses at the end of sentences, initials and ordinal numbers.

In the type-based stage, Kiss and Strunk (2006) use the facts that (1) there is a strong collocational dependency as abbreviations (almost) always occur with periods, (2) abbreviations tend to be short and (3) often contain additional internal periods.

Next, they apply orthographic and collocation heuristics in the token-based stage. The fact that the sentence boundary is usually followed by a capitalized letter is cautiously used to figure out whether the boundary is preceeded by an abbreviation or ellipsis. As

abbreviations like `"Mr."` are almost always followed by a capitalized name, this approach is extended by counting how often the word appears with upper- and lowercase letter. A collocation around a period exists when two words appear almost always together, seperated by a period. Kiss and Strunk (2006) conclude that these collocations should be treated as abbreviations. NLTK contains a pre-trained Punkt tokenizer, as it requires huge amounts of training data. In the following, this pre-trained Punkt tokenizer is used to tokenize the text material.

### 3.2.3 Chunking

The article text is now broken down into words. As the identification of companies is needed in a further processing step, chunking is considered to treat company names like *Red Hat* as one instead of two expressions. To handle this task, all possible bigrams in the text are identified. A bigram is defined as a fixed two-word phrase like *New York* (Manning, 1999). The function `chunk_bigram` is used to solve this task and can be found in Listing 5.

```
def chunk_bigram(sentence):
        global chunk_list
        word_n     = len(sentence) - 1
        i          = 0
        while i < word_n:
                bigram_tmp = sentence[i] + ' ' + sentence[i+1]

                if bigram_tmp in chunk_list:
                        sentence[i] = sentence[i] + ' ' + sentence[i+1]
                        del sentence[i+1]
                        word_n      = len(sentence) - 1
                i = i + 1
        return(sentence)
```

**Listing 5:** Source Code to identify Company Bigrams TXTProcessing.py

In future work, the identification bigrams corresponding to company names can be extended to the identification of fixed expressions like *low volatility* or *high volatility*. This could lead to a better identification of positive and negative sentiment as it often depends on the specifying adjective how a noun should be perceived. However, the LM word list only contains single word expressions and the creation of a more sophisticated sentiment lexicon would go beyond the scope of this thesis but might be part of future work. Thus,

the chunking of specific adjective-noun combinations is not necessary.

### 3.2.4 Slicing

Up to now, the article is splitted into sentences and words. Also, multi-word company names are treated as a single word. However, it would still only be possible to count positive and negative sentiment words on article level. Wang et al. (2014) state that it would be too simplifying to assume one sentiment value for long articles that mention multiple stocks. The reason is that the sentiment for each stock may be different. A simple way to solve this issue would be to identify company names in each sentence, count sentiment words in sentences with company name and aggregate the values with respect to each company. But this would lead to lower sentiment values as sentiment words may appear in sentences without company name.

In the following, the distance based approach, also used in Wang et al. (2014), is followed for S&P 500 companies. One advantage of NASDAQ as data source is that the mentioned companies are already known for each article. Hence, the usage of NLP techniques such as Named Entity Recognition (NER) is not necessarily needed and the company names and stock symbols can be identified by a mapping list. This list obviously also contains the bigrams, a sequence consisting out of two words, for companies mentioned in Section 3.2.3. In the first step, sentences that explicitly mention a company or a stock symbol ar identified and tagged with the relevant symbol. The tagged sentences are used as landmarks in the further slicing process and may contain more than one stock symbol. The stocks that are possibly identified are further limited to the tagged stock symbols in the meta-data as other company names like *Facebook* might appear without further relevance in the article.

The corresponding Python code to identify the landmarks is contained in Listing 6. The needed input for this function are on the one hand `sentence`, the sentence that should be tagged, and on the other hand `symbols`, the stock symbols that appear in the article's meta data. Additionally, a globally available company dictionary `comp_dict` must be specified that allows to map expressions to specific companies. Secondly, sentences before the first and after the last landmark get these landmarks assigned, respectively. Thirdly, sentences between two landmarks are tagged with the closest stock symbol. However, it is also checked whether the article is a priori tagged with the stock symbols to ensure that company names which may appear without being part of the article discussion, e.g. *Facebook*, are not assigned.

```
def find_lm(sentence, symbols):
global comp_dict
comp      = []
[comp.append( comp_dict[word] ) for word in sentence if word in comp_dict]
# Return only unique stock symbols
comp      = sorted(set(comp))
comp      = [symbol for symbol in comp if symbol in symbols]
return(comp)
```

**Listing 6:** Source Code to identify Landmarks in Sentence TXTProcessing.py

As some stock symbols are ambiguous, inconsiderately identifying stock symbols by a mapping table could lead to problems. Examples of this ambiguity are the symbol "A" for *Agilent Technologies Inc.*, "GAS" for *AGL Resources Inc.* and "JOY" for *Joy Global Inc.* While the problems are less severe for "JOY" as it probably really refers to *Joy Global Inc.* if it appears in an article that is a priori marked with "JOY", it is definitely problematic for *Agilent Technologies Inc.* as "A" is a frequent word in the English language. Also, "GAS" could probably also appear in sentences about oil and gas companies and refer to the natural resource instead of the company. To reduce misclassification errors, the stock symbols "A" and "GAS" are not used in the tagging process while (part) of the company names, namely "Agilent" and "AGL Resources" are used.

It must be stated that there are more refined ways to assign landmarks, e.g. using additional entities that are close the specific companies like names of the chief executive officers, product names or even the location of the company's headquarters. These techniques are collected under the term Named Entity Recognition (NER). Nadeau and Sekine (2007) provide a survey of NER and state that both rule and machine learning based approaches are highly domain dependent. Here, the domain is business and economics and more specifically companies in the S&P 500. Machine learning techniques would require huge amounts of already tagged training data which is currently, at least for the public, not available.

### 3.3 Sentiment Tagging

The sentiment tagging itself is straight forward. Every word in the article is compared to the LM lexicon to decide whether it should be treated as positive, negative or neutral. The approach of Hu and Liu (2004) and Zhang et al. (2015) is closely followed to handle negation in the text. The words that indicate negation are "not", "never", "no", "neither",

"nor", "none" and "n't". The polarity of a positive (negative) word is switched to negative (positive) if there is one of the named negation words in close proximity. As in Hu and Liu (2004) and Zhang et al. (2015), this is the case if the distance of the word and negation phrase is less or equal to 5 words. Additionally to this approach, only negation words are considered that are inside the sentence boundaries.

| Variable | Description |
|---|---|
| $W_{i,j,t}$ | Total number of words in article $i$ about company $j$ on day $t$ |
| $W^{+}_{i,j,t}$ | Number of positive words in article $i$ about company $j$ on day $t$ |
| $W^{-}_{i,j,t}$ | Number of negative words in article $i$ about company $j$ on day $t$ |
| $S_{i,j,t}$ | Total number of sentences in article $i$ about company $j$ on day $t$ |
| $S^{+}_{i,j,t}$ | Number of positive sentences in article $i$ about company $j$ on day $t$ |
| $S^{-}_{i,j,t}$ | Number of negative sentences in article $i$ about company $j$ on day $t$ |

**Table 3:** Derived Sentiment Variables for Article $i$ about Company $j$ on Day $t$

Next, sentiment on both word and sentence level is considered and the corresponding sentiment variables are derived. These variables are summarized in Table 3 with $i$, $j$ and $t$ referring to article, company and day, respectively. Further, the variables are aggregated to obtain a signal on day level instead of article level. Here, Antweiler and Frank (2004) consider a measure of bullishness while Chen et al. (2014) and Zhang et al. (2015) use the proportion of negative and positive words for each day. The fraction of positive words about company $j$ on day $t$ is given as

$$P^{W}_{j,t} = \frac{\sum_{i=1}^{I_t} W^{+}_{i,j,t}}{\sum_{i=1}^{I_t} \mathbf{1}(W^{Pol}_{i,j,t} > 0)\, W_{i,j,t}} \tag{1}$$

where $I_t$ denotes the total number of articles on day $t$ and $W^{Pol}_{i,j,t} = W^{+}_{i,j,t} + W^{-}_{i,j,t}$ as number of words with positive polarity in article $i$ for company $j$ on day $t$. The fraction of negative words $N^{W}_{j,t}$, positive sentences $P^{S}_{j,t}$ and negative sentences $N^{S}_{j,t}$ for company $j$ on day $t$ are calculated analogously.

Furthermore, the indicator whether there is at least one article about company $j$ on day $t$ is given as

$$Ind_{j,t} = \begin{cases} 1, & \text{if } I_t > 0. \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Define the word based measure of bullishness for company $j$ on day $t$ as

$$B_{j,t}^{W} = \log(2)^{-1} \log\left\{(1 + P_{j,t}^{W}) \,/\, (1 + N_{j,t}^{W})\right\} \tag{3}$$

and accordingly the sentence based measure of bullishness as

$$B_{j,t}^{S} = \log(2)^{-1} \log\left\{(1 + P_{j,t}^{S}) \,/\, (1 + N_{j,t}^{S})\right\}. \tag{4}$$

It can easily be seen that $B_{j,t}^{W} \in [-1,\, 1]$ and $B_{j,t}^{S} \in [-1,\, 1]$ which improves the interpretability of these measures.

However, as the length of an article might affect the quality of the derived measure, we also consider the modified versions of the previous bullishness measures. The modified bullishness measure for company $j$ on day $t$ is given by a simple weighting scheme:

$$B_{j,t}^{W*} = \log(W_{j,t})\, B_{j,t}^{W} \quad \text{and} \quad B_{j,t}^{S*} = \log(S_{j,t})\, B_{j,t}^{S}. \tag{5}$$

One problem of this bullishness measure might be that positive and negative sentiment would have an opposing effect in a linear regression model. While this might be appropriate in case of returns as dependent variable, it is clearly not sufficient in case of volatility or trading volume. Hence, define the functions $Neg(\cdot)$ and $Pos(\cdot)$ that refer to

$$Neg(x) = \begin{cases} x, & \text{if } x < 0 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

Thus, we are able to check whether the effects of sentiment are asymmetric or not. Note that the indices of each measure might be dropped if it is clear which measure being referred to.

# 4 Data

In this chapter, the main sources of data for the empirical analysis are introduced.

## 4.1 Financial Articles

In total, 164,148 financial articles are scraped from the NASDAQ page and loaded into a MySQL database by using the techniques described in Chapter 2. However, many of these articles are not about S&P 500 companies as there are also articles about other stock markets, commodities and even about general investing without naming a specific product. 79,231 articles about S&P 500 firms as of October 2014 remain. Figure 4 shows

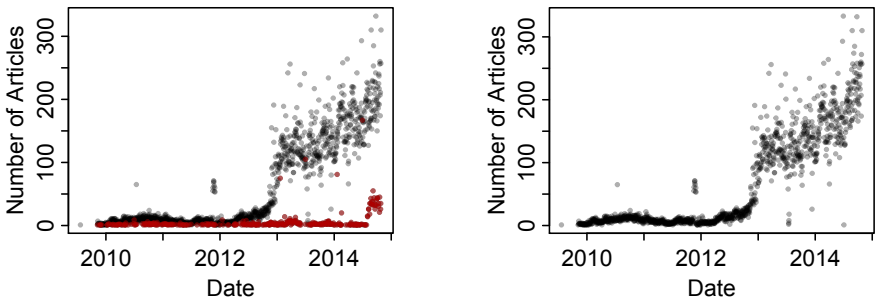


**Figure 4:** Number of S&P 500 Articles per Day before Shifting (left Panel) and after Shifting (right Panel); red indicates Days with closed Stock Markets and early Closing of Stock Markets. Shifting.R

the number of S&P 500 articles over time. Two distinct features can be observed: The number of articles per day increases over time and there are days on which hardly any articles are published. The increase of articles could either be the result of a larger number of contributors or the existing contributors are publishing content more frequently. However, the answer to this question is postponed. The question, why there seem to be days with less articles can be answered by looking at the days on which the U.S. stock markets are closed. These dates are colored red in Figure 4 and it is easily observable that the closing days match days with less articles. Furthermore, simply disregarding this fact would lead to discarding the signals of these articles as there are no observable stock reactions on this day due to the closed markets. Hence, the signals of articles on closing can either be excluded or included by shifting the articles to the last day with trading. This does not lead to a look-ahead bias as only the lagged sentiment values are used to model the stock reactions. We can assume that new information about companies is published as soon as

it becomes public. Hence, the articles on days with closed stock markets are shifted and included in the further analysis. The number of shifted S&P 500 articles over time can be seen in the right panel of Figure 4.

It is also observable in Figure 4 that there seem to be days without articles before November 2014. A look at the data shows that the articles are published between 2009-07-21 and 2014-10-25 but only four articles are published before November 2009. These four articles are excluded. The language processing steps described in Section 3.2 are applied and $162,237$ observations of sentiment on firm level are obtained. Some of the observations do not include a sentiment polarity on sentence level that is different from zero and more specifically that $P^S_{i,j,t} \neq N^S_{i,j,t}$ holds. A situation with $P^S_{i,j,t} = N^S_{i,j,t}$ would correspond to a neutral sentiment score for article $i$ about company $j$ on day $t$. Antweiler and Frank (2004) have already concluded that neutral sentiment just lowers the signal to noise ratio and thus, observations with neutral sentiment are excluded. This step leads to $71,708$ remaining articles and $123,544$ observations.

Looking at the frequencies of stock symbols per articles in Table 4.1 leads to another interesting finding. Only 8% of the articles that include at least one S&P 500 company

| S&P \ All | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | ≥13 | ∑ | % |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 6,012 | 2,834 | 3,861 | 10,405 | 4,737 | 191 | 121 | 240 | 27 | 60 | 19 | 56 | 17 | 28,580 | 40 |
| 2 | | 1,535 | 6,045 | 8,983 | 5,072 | 412 | 303 | 543 | 30 | 40 | 17 | 80 | 11 | 23,071 | 32 |
| 3 | | | 1,214 | 5,744 | 4,126 | 238 | 233 | 596 | 32 | 32 | 23 | 103 | 19 | 12,360 | 17 |
| 4 | | | | 2,564 | 2,640 | 39 | 26 | 21 | 16 | 23 | 26 | 101 | 14 | 5,470 | 8 |
| 5 | | | | | 1,298 | 31 | 18 | 17 | 14 | 22 | 23 | 111 | 17 | 1,551 | 2 |
| 6 | | | | | | 15 | 13 | 13 | 12 | 24 | 18 | 68 | 14 | 177 | < 1 |
| 7 | | | | | | | 9 | 12 | 12 | 10 | 12 | 66 | 12 | 133 | < 1 |
| 8 | | | | | | | | 0 | 10 | 21 | 6 | 47 | 10 | 94 | < 1 |
| 9 | | | | | | | | | 3 | 20 | 4 | 59 | 12 | 98 | < 1 |
| 10 | | | | | | | | | | 17 | 4 | 53 | 9 | 83 | < 1 |
| 11 | | | | | | | | | | | 1 | 28 | 16 | 45 | < 1 |
| 12 | | | | | | | | | | | | 20 | 8 | 28 | < 1 |
| ≥13 | | | | | | | | | | | | | 18 | 18 | < 1 |
| ∑ | 6,012 | 4,369 | 11,120 | 27,696 | 17,873 | 926 | 723 | 1,442 | 156 | 269 | 153 | 792 | 177 | 71,708 | 100 |
| % | 8 | 6 | 16 | 39 | 25 | 1 | 1 | 2 | < 1 | < 1 | < 1 | 1 | < 1 | 100 | |

**Table 4:** Frequencies of all Stock Symbols and Stock Symbols belonging to S&P 500 Firms in Articles about at least one S&P 500 Firm TXTSummary.R

articles about one single company. While the articles are sliced according to Section 3.2.4 to avoid multicollinearity and noise in the sentiment signals, the simple distance based

slicing process itself is probably not robust if there are too many mentioned stock symbols in one article. There is no previous research about the critical number of companies per article, so an intuitive decision is needed to proceed. One article actually contains 99 stock symbols and here, a proper distinction of sentiment on firm level is clearly not possible. More than 99% of the articles are about less than 13 companies and this seems to be a reasonable limit for the number of companies per article. Accordingly, articles that mention at least 13 companies are excluded.

After the previous cleaning steps, 71,531 articles and 122,607 observations on firm level remain. We can now take a closer look at the summary statistics for the articles in Table 5. The upper part of Table 5 contains statistics for the articles on author level, sorted by the total number of published articles. As there are 101 different contributors, the contributors with less than 400 published articles are aggregated into the group *Various*. Roughly 95 % of the articles have been created by the other 15 contributors. The previous question, whether there are more articles per contributor or simply more contributors that lead to the increased amount of articles in more recent years is now answered by that.

There are several new publishers after 2012 like *BNK Invest* or *Minyanville*. However, we can also see by looking at the Median of the date variable that publishers like *Zacks.com* and *Nasdaq.com* published roughly half of their content in the last year. Hence, there are both more publishers and the already existing publishers started to use the platform more frequently in the last years. This leads to the conclusion that the NASDAQ article platform might be even more important in the future as it collects articles of different sources and seems to be accepted by other well known platforms as only *David Sterman* stopped publishing in 2013.

There is no significant difference of symbols per articles for the different contributors. However, the average number of sentences with clear negative or positive polarity $\mu_{Pol}$ differs for the contributors. Nonetheless, this might be due to different article lengths as the overall length also differs. For example articles by *Seeking Alpha* are normally much longer than articles by *MT Newswires* which tend to be quite short.

A direct comparison of the data set to other data sets in the literature is not really possible but Antweiler and Frank (2004) state that there are most frequently between 20 and 50 words in a Yahoo stock board message. While they obtained with 1.5 million messages far more messages than the NASDAQ sample contains, it is also clear that these messages are quite short. If we assume that the standard Yahoo message includes 35

words then they have overall 5.25 million words while the NASDAQ set about S&P 500 companies contains more than 67.5 million words in 3.4 million sentences. Obviously, the NASDAQ set is quite large in comparison and might lead to additional results in comparison to other data sources.

The lower panel of Table 5 refers to observations aggregated on GICS sector level. Note that these observations do not necessarily correspond to the number of articles as there are articles including companies in different sectors. The sectors *Information Technology* and *Consumer Discretionary* contain most of the observations while there are the least observations for *Telecommunication Services*. However, the pattern of publishing over time and language statistics for the observations are similar in each sector. Hence, we can assume that contributors publish balanced in all sectors.

### 4.2 Stock Data

Stock specific data is collected from Datastream and Compustat. Datastream is used to gather the S&P 500 constituent list of October 2014. Furthermore, daily prices and trading volume are collected for these constituents. The daily trading volume is defined as number shares traded on a day. Compustat is used to gather Global Industry Classification Standard (GICS) sector for these assets. In the following, three stock reactions are considered: volatility, detrended log trading volume and return.

Due to the observations on day-level, we are interested in a measure of volatility that captures the variability of the stock price over a day. Such a measure, the realized volatility, can be obtained by using high-frequency intra-day returns. Garman and Klass (1980) show that this estimator might be improved by using high-low data and define the range-based measure of volatility for company $j$ on day $t$ as

$$\begin{aligned} \sigma_{j,t} &= 0.511(u-d)^2 - 0.019\left\{c(u+d) - 2ud\right\} - 0.838c^2 \qquad (7) \\ \text{with } u &= \log(P_{j,t}^H) - \log(P_{j,t}^L), \\ d &= \log(P_{j,t}^L) - \log(P_{j,t}^O), \\ c &= \log(P_{j,t}^C) - \log(P_{j,t}^O), \end{aligned}$$

where $P_{j,t}^H$, $P_{j,t}^L$, $P_{j,t}^O$, $P_{j,t}^C$ are the daily highest, lowest, opening and closing stock prices, respectively.

It is shown by Chen et al. (2006) and Shu and Zhang (2006) that the Garman and Klass range-based measure of volatility provides equivalent results to the realized volatility

| | Date | | | | Articles | | | | Sentences | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | Min | Q2 | Max | $\mu_{SP500}$ | $\sigma_{SP500}$ | $\mu_{Pol}$ | $\sigma_{Pol}$ | Min | Q1 | Q2 | Q3 | Max | $\mu_{Words}$ | $\sigma_{Words}$ |
| Contributors | | | | | | | | | | | | | | | |
| Zacks.com | 33,045 | 2011-11-03 | 2013-12-20 | 2014-10-28 | 1.95 | 1.01 | 11.09 | 10.11 | 1 | 20 | 32 | 52 | 648 | 20.86 | 6.63 |
| NASDAQ.com News | 12,848 | 2010-06-18 | 2013-10-29 | 2014-10-28 | 2.14 | 1.29 | 3.05 | 2.29 | 5 | 14 | 18 | 24 | 828 | 16.12 | 1.43 |
| MT Newswires | 3,591 | 2010-02-23 | 2014-04-21 | 2014-10-28 | 1.45 | 0.8 | 6.7 | 10.75 | 2 | 8 | 11 | 16 | 155 | 24.2 | 8.37 |
| Various | 3,408 | 2009-07-22 | 2012-03-15 | 2014-10-28 | 2.06 | 1.49 | 14.4 | 14.12 | 1 | 18 | 38 | 72 | 1,044 | 23.58 | 10.35 |
| Motley Fool | 3,396 | 2010-08-24 | 2014-09-19 | 2014-10-28 | 1.97 | 1.06 | 19.12 | 8.58 | 18 | 49 | 66 | 118 | 376 | 18.98 | 2.42 |
| BNK Invest | 3,106 | 2014-01-09 | 2014-06-11 | 2014-10-28 | 2.06 | 1.14 | 5.16 | 2.87 | 3 | 9 | 14 | 24 | 78 | 26.3 | 6.6 |
| Benzinga | 2,599 | 2012-04-03 | 2013-04-18 | 2014-10-27 | 2.07 | 1.17 | 11.37 | 6.61 | 4 | 30 | 42 | 70 | 355 | 18.19 | 7.28 |
| Minyanville | 2,141 | 2013-04-16 | 2013-08-14 | 2014-10-27 | 2.5 | 1.23 | 13.99 | 13.58 | 3 | 28 | 52 | 86 | 1,056 | 21.17 | 5.92 |
| GuruFocus | 1,817 | 2012-04-30 | 2013-05-21 | 2014-10-24 | 2.31 | 1.29 | 11.08 | 12.59 | 4 | 40 | 70 | 126 | 1,195 | 18.29 | 3.36 |
| Schaeffer's IR | 1,132 | 2010-04-30 | 2012-04-30 | 2014-06-19 | 2.94 | 1.37 | 17.33 | 10.92 | 15 | 43 | 88 | 129 | 285 | 20.47 | 1.93 |
| Investor's BD | 1,063 | 2012-06-21 | 2013-09-27 | 2014-10-28 | 2.15 | 1.18 | 16.65 | 14.6 | 13 | 27 | 54 | 114 | 915 | 17.52 | 2.9 |
| SeekingAlpha | 1,042 | 2009-11-06 | 2012-01-29 | 2014-10-28 | 2.15 | 1.28 | 38.08 | 32.72 | 4 | 59 | 108 | 208 | 1,542 | 21.8 | 4.67 |
| David Sterman | 791 | 2010-02-12 | 2011-08-26 | 2013-04-30 | 2.28 | 1.31 | 20.59 | 12.3 | 25 | 47 | 78 | 122 | 476 | 20.08 | 1.77 |
| StreetAuthority | 674 | 2009-11-17 | 2013-05-01 | 2014-10-24 | 2.09 | 1.31 | 18.89 | 10.49 | 13 | 40 | 58 | 105 | 651 | 19.65 | 3.02 |
| Wyatt IR | 473 | 2010-08-09 | 2013-02-01 | 2014-07-08 | 1.99 | 1.2 | 13.04 | 10.62 | 2 | 27 | 41 | 74 | 420 | 16.95 | 3.14 |
| Kapitall | 404 | 2010-11-15 | 2011-09-17 | 2014-09-11 | 3.53 | 2.69 | 10.5 | 6.51 | 12 | 50 | 73 | 176 | 1,890 | 15.98 | 5.32 |
| Sectors | | | | | | | | | | | | | | | |
| Consumer Discretionary | 16,195 | 2009-11-06 | 2013-11-21 | 2014-10-28 | 2.31 | 1.34 | 9.27 | 9.88 | 1 | 18 | 30 | 50 | 714 | 20.44 | 7.13 |
| Consumer Staples | 7,097 | 2009-11-11 | 2013-11-19 | 2014-10-28 | 2.53 | 1.46 | 8.5 | 9.34 | 1 | 17 | 28 | 48 | 666 | 20.72 | 7.57 |
| Energy | 5,701 | 2009-11-11 | 2013-12-09 | 2014-10-28 | 2.14 | 1.47 | 8.42 | 9.39 | 1 | 15 | 26 | 48 | 658 | 21.51 | 9.1 |
| Financials | 14,713 | 2009-11-10 | 2013-11-15 | 2014-10-28 | 2.27 | 1.33 | 8.02 | 10.43 | 1 | 18 | 24 | 42 | 756 | 19.49 | 6.44 |
| Health Care | 9,315 | 2009-11-09 | 2014-01-23 | 2014-10-28 | 2.23 | 1.43 | 8.62 | 9.73 | 1 | 15 | 26 | 46 | 1,053 | 20.38 | 7.03 |
| Industrials | 9,265 | 2009-07-22 | 2013-11-25 | 2014-10-28 | 2.21 | 1.45 | 7.47 | 8.47 | 1 | 14 | 23 | 42 | 840 | 19.94 | 6.24 |
| Information Technology | 16,861 | 2009-11-10 | 2013-11-06 | 2014-10-28 | 2.42 | 1.38 | 10.12 | 11.4 | 1 | 19 | 33 | 56 | 1,542 | 20.54 | 6.82 |
| Materials | 3,957 | 2009-07-22 | 2013-10-22 | 2014-10-28 | 2.1 | 1.47 | 8.12 | 9.35 | 1 | 16 | 25 | 42 | 531 | 20.42 | 7.17 |
| Telecommunication Services | 2,086 | 2009-11-06 | 2013-10-24 | 2014-10-28 | 2.65 | 1.54 | 7.51 | 8.29 | 1 | 17 | 28 | 45 | 285 | 20.64 | 6.05 |
| Utilities | 2,374 | 2009-12-02 | 2013-11-25 | 2014-10-28 | 2.44 | 1.5 | 7.78 | 9.53 | 1 | 17 | 27 | 52 | 500 | 20.43 | 8.84 |

Q1, Q2 and Q3 represent 25%, 50% and 75% quantile, respectively. $\mu_{SP500}$ and $\sigma_{SP500}$ denote average number and standard deviation of S&P 500 firms per article, respectively. $\mu_{Pol}$ and $\sigma_{Pol}$ denote average and standard deviation of sentences with polarity (either positive or negative) per article, respectively. $\mu_{Words}$ and $\sigma_{Words}$ denote the average and standard deviation of words per sentence.

**Table 5:** Summary Statistics for all S&P 500 Articles by Contributor and Sector TXTSummary.R

on daily level. Subsequently, the Garman and Klass range-based measure of volatility is used in the further analysis.

Following Girard and Biswas (2007), the detrended log trading volume for each stock is estimated by using a quadratic time trend equation:

$$V_{j,t}^{*} = \alpha + \beta_1 t + \beta_2 t^2 + V_{j,t}, \tag{8}$$

where $V_{i,t}^{*}$ corresponds to the raw daily log trading volume and the detrended log trading volume $V_{i,t}$ are the residuals. A look-ahead bias is avoided by using a rolling window of 120 observations and estimating a one-step ahead pseudo out-of-sample forecast.

Furthermore, the returns are calculated as

$$R_{j,t} = \log(P_{j,t}^{C}) - \log(P_{j,t-1}^{C}). \tag{9}$$

# 5 Empirical Results

The empirical results are presented in this section.

## 5.1 Contemporaneous Regression

Following Antweiler and Frank (2004), the effects of sentiment on the stock reactions ()volatility, trading volume and returns) are investigated by using contemporaneous regressions. Since Fama (1970), the EMH is widely accepted and leads to the assumption that news spreads quickly and is directly incorporated in stock prices and thus, the other mentioned stock reactions. Following, stock prices fully reflect all available information at each time point $t$. However, since the first mention of the efficient market hypothesis, it has been shown that real markets are not always efficient and due to this fact, stock prices are at least partially predictable as stated in Malkiel (2003). Nonetheless, the sentiment of news should have a significant impact on the stock reactions on the day the news arises. As the data is aggregated on a daily level we can not say whether stock reactions lead to specific news or whether sentiment in news influences the nature of the stock reactions.

In this section, panel regression models with fixed effects for each company are estimated. The models are given by

$$\sigma_{j,t} = \alpha_j + \beta_1^\top Sent_{j,t} + \beta_2^\top X_{j,t} + \gamma_j + \varepsilon_{j,t}, \tag{10}$$

$$V_{i,t} = \alpha_j + \beta_1^\top Sent_{j,t} + \beta_2^\top X_{j,t} + \gamma_j + \varepsilon_{j,t}, \tag{11}$$

$$R_{i,t} = \alpha_j + \beta_1^\top Sent_{j,t} + \beta_2^\top X_{j,t} + \gamma_j + \varepsilon_{j,t}. \tag{12}$$

As in Zhang et al. (2015), the models are estimated separately. $\gamma_j$ corresponds to the fixed effect for firm $j$ satisfying $\sum_j \gamma_j = 0$ and $\varepsilon_{j,t}$ is the error term of company $j$ at day $t$. Recall that several measures of sentiment have been derived in Section 3.3. Different versions of $Sent_{j,t}$ are considered, depending on the set of sentiment measures. Model 1 uses $Ind_{j,t}$, $P_{j,t}^W$ and $N_{j,t}^W$ as set of sentiment values on word level as well as $Ind_{j,t}$, $P_{j,t}^S$ and $N_{j,t}^S$ on sentence level. Model 2 and 3 incorporate the derived bullishness measures. More specifically, the set of Model 2 consists of $B_{j,t}^W$ and $Neg(B_{j,t}^W)$ on word level and $B_{j,t}^S$ and $Neg(B_{j,t}^S)$ on sentence level.

The Model 3 set contains $B_{j,t}^{W*}$ and $Neg(B_{j,t}^{W*})$ on word level as well as $B_{j,t}^{S*}$ and $Neg(B_{j,t}^{S*})$ on sentence level. Since Model 2 and Model 3 might not be easily interpretable regarding the dependent variabes $\sigma_{j,t}$ and $V_{i,t}$, Model 4 and Model 5 are adjusted such that the absolute value of the bullishness measure is included.

| Variable | Model 1 | | Model 2 | | Model 3 | |
|---|---|---|---|---|---|---|
| | Panel A: Volatility $\sigma_{j,t}$ | | | | | |
| $R_{j,t}$ | −0.034*** | (0.001) | −0.034*** | (0.001) | −0.034*** | (0.001) |
| $R_{M,t}$ | 0.008*** | (0.002) | 0.008*** | (0.002) | 0.008*** | (0.002) |
| $VIX_t$ | 0.000*** | (0.000) | 0.000*** | (0.000) | 0.000*** | (0.000) |
| $Ind_{j,t}$ | 0.000** | (0.000) | | | | |
| $N_{j,t}$ | 0.006** | (0.003) | | | | |
| $P_{j,t}$ | −0.002 | (0.003) | | | | |
| $B_{j,t}$ | | | 0.003 | (0.002) | | |
| $Neg(B_{j,t})$ | | | −0.011*** | (0.003) | | |
| $B^*_{j,t}$ | | | | | 0.001* | (0.000) |
| $Neg(B^*_{j,t})$ | | | | | −0.002*** | (0.001) |
| | Panel B: Detrended Log Trading Volume $V_{j,t}$ | | | | | |
| $\sigma_{j,t}$ | 7.816*** | (0.221) | 7.906*** | (0.221) | 7.890*** | (0.221) |
| $R_{j,t}$ | 0.969*** | (0.075) | 0.979*** | (0.075) | 0.987*** | (0.075) |
| $R_{M,t}$ | −3.396*** | (0.136) | −3.490*** | (0.137) | −3.482*** | (0.137) |
| $VIX_t$ | 0.001*** | (0.000) | 0.000 | (0.000) | 0.000** | (0.000) |
| $Ind_{j,t}$ | 0.043*** | (0.005) | | | | |
| $N_{j,t}$ | 3.380*** | (0.205) | | | | |
| $P_{j,t}$ | 0.281 | (0.248) | | | | |
| $B_{j,t}$ | | | 1.036*** | (0.143) | | |
| $Neg(B_{j,t})$ | | | −4.195*** | (0.224) | | |
| $B^*_{j,t}$ | | | | | 0.232*** | (0.027) |
| $Neg(B^*_{j,t})$ | | | | | −0.898*** | (0.042) |
| | Panel C: Returns $R_{j,t}$ | | | | | |
| $\sigma_{j,t}$ | −0.294*** | (0.008) | −0.294*** | (0.008) | −0.294*** | (0.008) |
| $R_{M,t}$ | 1.040*** | (0.004) | 1.040*** | (0.004) | 1.040*** | (0.004) |
| $VIX_t$ | −0.000 | (0.000) | −0.000 | (0.000) | −0.000 | (0.000) |
| $Ind_{j,t}$ | −0.000 | (0.000) | | | | |
| $N_{j,t}$ | −0.025*** | (0.008) | | | | |
| $P_{j,t}$ | 0.044*** | (0.009) | | | | |
| $B_{j,t}$ | | | 0.016*** | (0.005) | | |
| $Neg(B_{j,t})$ | | | 0.016* | (0.008) | | |
| $B^*_{j,t}$ | | | | | 0.003*** | (0.001) |
| $Neg(B^*_{j,t})$ | | | | | 0.004*** | (0.002) |

*** refers to a $p$ value less than 0.01, ** refers to a $p$ value more than or equal to 0.01 and smaller than 0.05, and * refers to a $p$ value more than or equal to 0.05 and less than 0.1. Values in parentheses are standard errors.

**Table 6:** Contemporaneous Regression Results on Word Level

TXTPanelContemp.R

$X_{j,t}$ is a vector of variables to control for systematic risk that always includes (1) S&P 500 index return ($R_{M,t}$) to control for general market returns and (2) the CBOE VIX index on date $t$ to measure the generalized risk aversion ($VIX_t$). Furthermore, a set of firm idiosyncratic variables that differs according to the dependent variable is used. In equation 10 and 12 we include only $R_{i,t}$ or $\sigma_{j,t}$, respectively. Both $\sigma_{j,t}$ and $R_{i,t}$ are included in equation 11.

| Variable | Model 4 | | Model 5 | |
|---|---|---|---|---|
| | Panel A: Volatility $\sigma_{j,t}$ | | | |
| $\lvert B_{j,t}\rvert$ | 0.003 | (0.002) | | |
| $Neg(B_{j,t})$ | −0.005* | (0.003) | | |
| $\lvert B^*_{j,t}\rvert$ | | | 0.001* | (0.000) |
| $Neg(B^*_{j,t})$ | | | −0.001** | (0.000) |
| | Panel B: Detrended Log Trading Volume $V_{j,t}$ | | | |
| $\lvert B_{j,t}\rvert$ | 1.036*** | (0.143) | | |
| $Neg(B_{j,t})$ | −2.123*** | (0.203) | | |
| $\lvert B^*_{j,t}\rvert$ | | | 0.232*** | (0.027) |
| $Neg(B^*_{j,t})$ | | | −0.434*** | (0.037) |

*** refers to a $p$ value less than 0.01, ** refers to a $p$ value more than or equal to 0.01 and smaller than 0.05, and * refers to a $p$ value more than or equal to 0.05 and less than 0.1. Values in parentheses are standard errors.

**Table 7:** Contemporaneous Regression Results on Word Level: Model 4 and 5

TXTPanelContemp45.R

Table 6 shows the results of the contemporaneous regression on word level. The estimated parameters for the control variables do not change much across models and hence, their effect on the stock reactions appears to be quite robust. The parameter that accounts for negative sentiment is significant in all models with $\sigma_{j,t}$ as dependent variable while the effect of positive sentiment is only significant in Model 3. Furthermore, we can observe that the effect of $Neg(B^W_{j,t})$ and $Neg(B^{W*}_{j,t})$ is negative. Thus, more negative sentiment leads to a higher volatility in Model 2 as the absolute value for the parameter for a negative bullishness measure is much smaller than the parameter for a positive one. This is not as clear for Model 2 as increased negative sentiment has more or less the same effect as positive sentiment.

By taking a look at the estimated parameters for the detrended log trading volume $V_{j,t}$, the asymmetric response to sentiment becomes obvious. Again, the parameters for

negative sentiment, represented by $N_{j,t}$, $Neg(B_{j,t})$ and $Neg(B^*_{j,t})$ are highly significant. By checking the sign of the parameters, we can conclude that negative sentiment is connected to a higher trading volume. While positive sentiment also leads to an increased trading volume, the effect is not as large as for negative sentiment.

Results of Model 4 and Model 5 for volatility and trading volume can be found in Table 7. The results of course correspond to the previous regression but here, the asymmetry of sentiment can easily be seen. An increase of negative and positive sentiment co-occurs with larger values of volatility and trading volume. However, the increase is larger for negative sentiment than positive sentiment.

Panel C of Table 6 shows the regression results regarding the parameters with returns $R_{j,t}$ as dependent variable. Both negative and positive sentiment have a significant relationship with the returns of the same day but the nature of this relationship differs for the sentiment measures. While $|\beta(N_{j,t})| < \beta(P_{j,t})$, with $\beta(\cdot)$ referring to the estimated parameter of $\cdot$, indicates that the stock market participants might react stronger to positive than negative sentiment, the results differ for the bullishness measures. Since the parameters are significant and positive for $B_{j,t}$ and $B^*_{j,t}$, we can conclude that negative sentiment co-occurs with lower stock returns than positive sentiment. As the parameters for $Neg(B_{j,t})$ and $Neg(B^*_{j,t})$ are negative and significant, there seems to be an asymmetric response. As earlier stated, it cannot be said whether news arrive before the adjustment of the stock price. Hence, we cannot know whether the news pick up negative motion in the markets or the market reacts stronger to negative news.

The results for the derived sentiment values on sentence level are presented in Table 10 in the Appendix. While the results mostly correspond to the previously discussed results on word level, less parameters are significantly different from zero. Note that the smaller parameter values for the sentiment measures are due to the fact that the measures are often larger than before.

### 5.2 Time-Lagged Panel Regression

In this section, panel regression with lagged dependent variables is applied. The following panel regression models are estimated seperately for each stock reaction:

$$\sigma_{j,t+1} = \alpha_j + \beta_1^\top Sent_{j,t} + \beta_2^\top X_{j,t} + \gamma_j + \varepsilon_{j,t+1}, \quad (13)$$

$$V_{i,t+1} = \alpha_j + \beta_1^\top Sent_{j,t} + \beta_2^\top X_{j,t} + \gamma_j + \varepsilon_{j,t+1}, \quad (14)$$

$$R_{i,t+1} = \alpha_j + \beta_1^\top Sent_{j,t} + \beta_2^\top X_{j,t} + \gamma_j + \varepsilon_{j,t+1}. \quad (15)$$

Except for the usage of lagged variables, the same models and sentiment measures as in Section 5.1 are applied.

Table 8 shows the regression results for sentiment measures on word level. Regarding the future volatility $\sigma_{j,t+1}$, parameter estimates for the control variables are again stable over the different models. The estimated parameters for the sentiment measures in Model 1 are not significantly different from zero. However, the parameters that account for negative sentiment in Model 2 and 3 are both significant. As these parameters are negative, it can be concluded that negative sentiment on a day co-occurs with higher volatility on the next day. It can also be concluded that this effect is asymmetric as positive sentiment, covered by $B_{j,t}^W$ and $B_{j,t}^{W*}$, does not have a significant effect on the next day's volatility.

The regression results for the models with the detrended log trading volume $V_{j,t+1}$ as dependent variable are similar to the contemporaneous regression results. Here, the parameter values are smaller than before but still significantly different from zero. The interpretation remains similar as before: News in general lead to an increased trading volume. This effect is larger in size if there is more negative than positive sentiment in the news.

The hypothesis that there is an asymmetric response of $V_{j,t+1}$ is supported by the results in Table 8. While sentiment in news is generally followed by increased trading volume, the size of this effect is significantly larger for negative sentiment. However, this is not as clear for the volatility $\sigma_{j,t+1}$ as the estimated parameters are not significant. One appropriate way to further investigate this matter would be a simulation as in Zhang et al. (2015). As this is out of the scope of this thesis, this task is delayed to future work.

The results regarding the returns $R_{j,t+1}$ are more ambivalent. While at least one of the parameters that account for the effects of news is significant in each model, the signs of these parameters seem not reasonable. In Model 2, negative sentiment actually has

a positive effect on returns of the following day that corresponds in size to the effect of positive sentiment. Naturally, the negative sentiment in news should lead to decreased returns on the following day. Recall the results priorly discussed in Section 5.1. On a single day, negative sentiment is accompanied by lower returns. The returns are significantly more decreased than the increased returns that co-occur with positive sentiment. Hence, the question arises whether market participants are prone to overreact if there are negative news. Thus, the sign and size of the estimated parameters could correspond to the correction of a former overreaction. If this is the case, then the overreaction regarding news an interesting implication as the participants only seem to overreact if the news is bearish. This might correspond to the formerly established known risk averse behavior of investors. However, it definitely is in line with the UIH. Recall that in contrast to the EMH, the market participants set new prices before the full range of the news content is resolved. In case of both favorable and unfavorable news, the investors set stock prices significantly below their conditional expected values and thus, react risk-averse. As the uncertainty regarding the impact of news is resolved subsequently and the prices adjust to fully reflect the known information. Furthermore, the results correspond to prior findings by Brown et al. (1988) who extract "news events" from stock prices by looking for abnormally large returns and find a pattern of stock price adjustment that suggests the validity of the UIH. Hence, an extension of their empirical study is provided here as sentiment of real news articles is extracted. The additional results in Table 9 also point in the same direction. The stock price slightly revise up following both negative and positive news on the prior day. The additional effect of negative sentiment is not significant.

Again, the corresponding results for the sentiment measures on sentence level are presented in the Appendix. They can be found in Table 11. One again, the interpretation does not change.

# 6 Conclusion

In this thesis, articles are scraped from the internet and subsequently, sentiment is distilled. The utilized distillation process allows for sentiment measures on both word and sentence level. Also, a distance based entity disambiguation is performed.

To revise the research questions, take a look at question the question whether derived sentiment measure plays a role (1). The regression results point in the direction that a single measure of bullishness might provide advantages on a firm level investigation. Note that this result might differ from previous research due to the slicing process. However, While simple fractions lead to significant results in a contemporaneous regression model the estimated parameters are not significant in the time-lagged regression. This is not the case for the derived bullishness measures. A weighting scheme in the bullishness measure does not provide further insight. Furthermore, switching from word level sentiment to sentence level sentiment does not change the main results.

A possible asymmetric response is investigated in research question (2) for volatility and trading volume. The results point in the direction that these variables indeed behave differently depending on the fact whether the news is positive or negative. It can be concluded that these results are in line with previous research such as Zhang et al. (2015).

Thirdly, it is checked whether the results correspond to the uncertain information hypothesis. Positive news co-occur with positive returns while negative news co-occur with negative returns on the same day. However, this effect is more pronounced for negative news. On the next day, the prices adapt in a way that both negative and positive sentiment lead to increased returns. This is in line with the UIH but robustness checks should be performed in future work.

| Variable | Model 1 | | Model 2 | | Model 3 | |
|---|---|---|---|---|---|---|
| | Panel A: Volatility $\sigma_{j,t+1}$ | | | | | |
| $\sigma_{j,t-1}$ | 0.022*** | (0.003) | 0.022*** | (0.003) | 0.022*** | (0.003) |
| $R_{j,t}$ | −0.006*** | (0.001) | −0.006*** | (0.001) | −0.006*** | (0.001) |
| $R_{M,t}$ | −0.002 | (0.002) | −0.002 | (0.002) | −0.002 | (0.002) |
| $VIX_t$ | 0.000*** | (0.000) | 0.000*** | (0.000) | 0.000*** | (0.000) |
| $Ind_{j,t}$ | 0.000 | (0.000) | | | | |
| $N^W_{j,t}$ | 0.004 | (0.003) | | | | |
| $P^W_{j,t}$ | −0.002 | (0.003) | | | | |
| $B^W_{j,t}$ | | | 0.001 | (0.002) | | |
| $Neg(B^W_{j,t})$ | | | −0.006** | (0.003) | | |
| $B^{W*}_{j,t}$ | | | | | 0.000 | (0.000) |
| $Neg(B^{W*}_{j,t})$ | | | | | −0.001** | (0.001) |
| | Panel B: Detrended Log Trading Volume $V_{j,t+1}$ | | | | | |
| $\sigma_{j,t}$ | 3.571*** | (0.222) | 3.620*** | (0.222) | 3.609*** | (0.222) |
| $R_{j,t}$ | 0.190** | (0.075) | 0.193** | (0.075) | 0.198*** | (0.075) |
| $R_{M,t}$ | −3.642*** | (0.137) | −3.697*** | (0.137) | −3.692*** | (0.137) |
| $VIX_t$ | −0.000** | (0.000) | −0.001*** | (0.000) | −0.001*** | (0.000) |
| $Ind_{j,t}$ | 0.037*** | (0.005) | | | | |
| $N^W_{j,t}$ | 1.447*** | (0.206) | | | | |
| $P^W_{j,t}$ | −0.580** | (0.250) | | | | |
| $B^W_{j,t}$ | | | 0.356** | (0.144) | | |
| $Neg(B^W_{j,t})$ | | | −1.962*** | (0.225) | | |
| $B^{W*}_{j,t}$ | | | | | 0.095*** | (0.027) |
| $Neg(B^{W*}_{j,t})$ | | | | | −0.455*** | (0.042) |
| | Panel C: Returns $R_{j,t+1}$ | | | | | |
| $\sigma_{j,t}$ | −0.036*** | (0.010) | −0.036*** | (0.010) | −0.036*** | (0.010) |
| $R_{j,t}$ | −0.009*** | (0.003) | −0.009** | (0.003) | −0.009** | (0.003) |
| $R_{M,t}$ | −0.040*** | (0.006) | −0.041*** | (0.006) | −0.040*** | (0.006) |
| $VIX_t$ | 0.000*** | (0.000) | 0.000*** | (0.000) | 0.000*** | (0.000) |
| $Ind_{j,t}$ | 0.001*** | (0.000) | | | | |
| $N^W_{j,t}$ | −0.003 | (0.010) | | | | |
| $P^W_{j,t}$ | −0.001 | (0.012) | | | | |
| $B^W_{j,t}$ | | | 0.021*** | (0.007) | | |
| $Neg(B^W_{j,t})$ | | | −0.038*** | (0.010) | | |
| $B^{W*}_{j,t}$ | | | | | 0.004*** | (0.001) |
| $Neg(B^{W*}_{j,t})$ | | | | | −0.007*** | (0.002) |

*** refers to a $p$ value less than 0.01, ** refers to a $p$ value more than or equal to 0.01 and smaller than 0.05, and * refers to a $p$ value more than or equal to 0.05 and less than 0.1. Values in parentheses are standard errors.

**Table 8:** Time-Lagged Panel Regression Results on Word Level

TXTPanelTimeLagged.R

| Variable | Model 4 | | Model 5 | |
|---|---|---|---|---|
| | Panel A: Volatility $\sigma_{j,t+1}$ | | | |
| $\lvert B_{j,t}\rvert$ | 0.001 | (0.002) | | |
| $Neg(B_{j,t})$ | −0.004 | (0.003) | | |
| $\lvert B^*_{j,t}\rvert$ | | | 0.000 | (0.000) |
| $Neg(B^*_{j,t})$ | | | −0.001 | (0.000) |
| | Panel B: Detrended Log Trading Volume $V_{j,t+1}$ | | | |
| $\lvert B_{j,t}\rvert$ | 0.356** | (0.144) | | |
| $Neg(B_{j,t})$ | −1.25 *** | (0.204) | | |
| $\lvert B^*_{j,t}\rvert$ | | | 0.095*** | (0.027) |
| $Neg(B^*_{j,t})$ | | | −0.265*** | (0.037) |
| | Panel C: Returns $R_{j,t+1}$ | | | |
| $\lvert B_{j,t}\rvert$ | 0.021*** | (0.007) | | |
| $Neg(B_{j,t})$ | 0.005 | (0.009) | | |
| $\lvert B^*_{j,t}\rvert$ | | | 0.004*** | (0.001) |
| $Neg(B^*_{j,t})$ | | | 0.001 | (0.002) |

*** refers to a $p$ value less than 0.01, ** refers to a $p$ value more than or equal to 0.01 and smaller than 0.05, and * refers to a $p$ value more than or equal to 0.05 and less than 0.1. Values in parentheses are standard errors.

**Table 9:** Time-Lagged Regression Results on Word Level: Model 4 and 5

TXTPanelTimeLagged45.R

# Appendices

## A Article Example

### 3 Disastrous Mistakes McDonald’s Should Regret

**By:** Motley Fool
**Posted:** 10/26/2014 9:00:00 AM
**Referenced Stocks:** MCD

Source: McDonald's.

Investors are quickly losing their appetite for **McDonald's** , with shares of the fast-food chain down roughly 6% in 2014 versus a 4% increase for the **S&P 500** . The company has struggled recently, with U.S. same-store sales and global comparable salesrespectively dropping 4.1% and 3.3% last quarter. And while CEO Don Thompson is confident in the company's ability to regain momentum in the U.S., customers and investors are not so sure.

The companythat has defined the quick service restaurant industry is battling to grow sales against newer restaurants that compete on perceived food quality rather than on price. And McDonald's appears to lack a cohesive strategy to fend off these challengers.

All that said, here are the three worst mistakes McDonald's has made.

**Giving up megacompetitor Chipotle**
McDonald's at one point owned morethan 90% of high-growth burrito maker **Chipotle** . After McDonald's provided much-needed capital and logistical support, the company sold Chipotle in 2006 to "focus on the core business." During its ownership, McDonald's helped Chipotle grow from a Colorado-based "mom-and-pop" chain to a multistate operation with more than 500 locations.

Neither Chipotle nor McDonald's want you to remember McDonald's' former ownership. Source: Chipotle.

Although McDonald's received $1.5 billion for the sale, that stake is now valued at $17 billion -- more than 1,000% higher than McDonald's exit price. And that's not even the worst part. Chipotle was a way for McDonald's to enter the fast-growing trend of healthy food; instead, the Golden Arches nurtured perhaps the biggest challenge to its business model and turned a subsidiary into a fierce competitor.

**Hate our huge menu? You're not alone**
McDonald's has an obesity problem that continues to get worse. And that's nothing to do with the food itself, but rather the huge menus that can now double as medieval fortification. For perspective, the chain's menu has grown 70% since 2007. And while more offerings might seem like a good thing, large menus result in slower service and more flare-ups between franchisees and the corporation.

Bloated menus raise inventory costs for smaller franchisees and lead to lower profit margins. The McDonald's corporate franchise fee is based upon sales instead of profits, making it a smaller concern for the company overall. In addition, remember that restaurant food is perishable ... even at McDonald's, regardless of what you read on the Internet. And for franchisees, waste means less profit and investment into their businesses.

For the end consumer, huge menus leads to a worse customer experience -- there's a reason we refer to QSRs as "fast food," but once you remove the fast from the equation the value proposition

refer to QSRs as "fast food," but once you remove the fast from the equation the value proposition falls substantially. A recent study from *QSR* magazine found the average drive-thru wait at McDonald's to be 3 minutes and 9.5 seconds, the longest wait time in at least 15 years. And that's the average - not the "can you please pull forward and wait" time.

**Remember Morgan Spurlock?**
Morgan Spurlock's groundbreaking film *Supersize Me* is over 10 years old. And while the film isn't immune from criticism, its effect on McDonald's should not be understated. The documentary put a face on growing concerns about food quality, nutrition, and health, and put McDonald's squarely in the crosshairs of this discussion Americans were having.

In the decade since, McDonald's has continued to struggle with perceptions regarding the quality of its food. In addition to selling perhaps the best way to play the trend toward healthier diets -- Chipotle -- the company appears unable to change its image as a junk food purveyor. As a nod to its poor reputation, the company recently added a section to its website to address perceived food quality and nutritional concerns.

**The future**
McDonald's has its work cut out for it. Not only are sales falling in the U.S., but the company is now experiencing problems abroad.

Thompson plans to right the ship by focusing on customizable burgers with a "Create Your Taste" program that is eerily similar to **Burger King** 's "Have It Your Way" campaign.

One thing is clear though, unless this program can simplify the menu and improve McDonald's poor food quality image, the fast-food chain will continue to have its lunch handed to it by Chipotle and other young and hungry upstarts looking to feast on their older rival's missteps.

**Holding McDonald's for its dividend? Check out these top dividend stocks instead**
The smartest investors know that dividend stocks simply crush their non-dividend paying counterparts over the long term. That's beyond dispute. They also know that a well-constructed dividend portfolio creates wealth steadily, while still allowing you to sleep like a baby. Knowing how valuable such a portfolio might be, our top analysts put together a report on a group of high-yielding stocks that should be in any income investor's portfolio. To see our free report on these stocks, just click here .

Figure 5: Example of real Article

The article 3 Disastrous Mistakes McDonald's Should Regret originally appeared on Fool.com.

*Jamal Carnette has no position in any stocks mentioned. The Motley Fool recommends Chipotle Mexican Grill and McDonald's. The Motley Fool owns shares of Chipotle Mexican Grill. Try any of our Foolish newsletter services free for 30 days . We Fools may not all hold the same opinions, but we all believe that considering a diverse range of insights makes us better investors. The Motley Fool has a disclosure policy .*

# B Additional Tables

| Variable | Model 1 | | Model 2 | | Model 3 | |
|---|---|---|---|---|---|---|
| | | | Panel A: Volatility $\sigma_{j,t}$ | | | |
| $R_{j,t}$ | −0.034*** | (0.001) | −0.034*** | (0.001) | −0.034*** | (0.001) |
| $R_{M,t}$ | 0.008*** | (0.002) | 0.008*** | (0.002) | 0.008*** | (0.002) |
| $VIX_t$ | 0.000*** | (0.000) | 0.000*** | (0.000) | 0.000*** | (0.000) |
| $Ind_{j,t}$ | 0.000** | (0.000) | | | | |
| $N_{j,t}$ | 0.000** | (0.000) | | | | |
| $P_{j,t}$ | −0.000 | (0.000) | | | | |
| $B_{j,t}$ | | | 0.000* | (0.000) | | |
| $Neg(B_{j,t})$ | | | −0.001*** | (0.000) | | |
| $B^*_{j,t}$ | | | | | 0.000** | (0.000) |
| $Neg(B^*_{j,t})$ | | | | | −0.000*** | (0.000) |
| | | | Panel B: Detrended Log Trading Volume $V_{j,t}$ | | | |
| $\sigma_{j,t}$ | 7.820*** | (0.221) | 7.912*** | (0.221) | 7.881*** | (0.221) |
| $R_{j,t}$ | 0.965*** | (0.075) | 0.974*** | (0.075) | 0.992*** | (0.075) |
| $R_{M,t}$ | −3.393*** | (0.136) | −3.487*** | (0.137) | −3.479*** | (0.137) |
| $VIX_t$ | 0.001*** | (0.000) | 0.000 | (0.000) | 0.001*** | (0.000) |
| $Ind_{j,t}$ | 0.046*** | (0.006) | | | | |
| $N_{j,t}$ | 0.227*** | (0.017) | | | | |
| $P_{j,t}$ | 0.025 | (0.018) | | | | |
| $B_{j,t}$ | | | 0.080*** | (0.011) | | |
| $Neg(B_{j,t})$ | | | −0.319*** | (0.018) | | |
| $B^*_{j,t}$ | | | | | 0.043*** | (0.004) |
| $Neg(B^*_{j,t})$ | | | | | −0.174*** | (0.007) |
| | | | Panel C: Returns $R_{j,t}$ | | | |
| $\sigma_{j,t}$ | −0.294*** | (0.008) | −0.294*** | (0.008) | −0.294*** | (0.008) |
| $R_{M,t}$ | 1.041*** | (0.004) | 1.040*** | (0.004) | 1.040*** | (0.004) |
| $VIX_t$ | −0.000 | (0.000) | −0.000 | (0.000) | −0.000 | (0.000) |
| $Ind_{j,t}$ | −0.000** | (0.000) | | | | |
| $N_{j,t}$ | −0.001 | (0.001) | | | | |
| $P_{j,t}$ | 0.003*** | (0.001) | | | | |
| $B_{j,t}$ | | | 0.001*** | (0.000) | | |
| $Neg(B_{j,t})$ | | | 0.001* | (0.001) | | |
| $B^*_{j,t}$ | | | | | 0.000* | (0.000) |
| $Neg(B^*_{j,t})$ | | | | | 0.001*** | (0.000) |

*** refers to a $p$ value less than 0.01, ** refers to a $p$ value more than or equal to 0.01 and smaller than 0.05, and * refers to a $p$ value more than or equal to 0.05 and less than 0.1. Values in parentheses are standard errors.

**Table 10:** Contemporaneous Regression Results on Sentence Level

TXTPanelContempt.R

| Variable | Model 1 | | Model 2 | | Model 3 | |
|---|---|---|---|---|---|---|
| | Panel A: Volatility $\sigma_{j,t}$ | | | | | |
| $\sigma_{j,t-1}$ | 0.022*** | (0.003) | 0.022*** | (0.003) | 0.022*** | (0.003) |
| $R_{j,t-1}$ | −0.006*** | (0.001) | −0.006*** | (0.001) | −0.006*** | (0.001) |
| $R_{M,t-1}$ | −0.002 | (0.002) | −0.002 | (0.002) | −0.002 | (0.002) |
| $VIX_{t-1}$ | 0.000*** | (0.000) | 0.000*** | (0.000) | 0.000*** | (0.000) |
| $Ind_{j,t-1}$ | 0.000 | (0.000) | | | | |
| $N^S_{j,t-1}$ | 0.000 | (0.000) | | | | |
| $P^S_{j,t-1}$ | −0.000 | (0.000) | | | | |
| $B^S_{j,t-1}$ | | | 0.000 | (0.000) | | |
| $Neg(B^S_{j,t-1})$ | | | −0.000** | (0.000) | | |
| $B^{S*}_{j,t-1}$ | | | | | 0.000 | (0.000) |
| $Neg(B^{S*}_{j,t-1})$ | | | | | −0.000*** | (0.000) |
| | Panel C: Detrended Log Trading Volume $V_{j,t}$ | | | | | |
| $\sigma_{j,t-1}$ | 3.574*** | (0.222) | 3.624*** | (0.222) | 3.602*** | (0.222) |
| $R_{j,t-1}$ | 0.188** | (0.075) | 0.191** | (0.075) | 0.202*** | (0.075) |
| $R_{M,t-1}$ | −3.642*** | (0.137) | −3.697*** | (0.137) | −3.689*** | (0.137) |
| $VIX_{t-1}$ | 0.000** | (0.000) | −0.001*** | (0.000) | −0.001*** | (0.000) |
| $Ind_{j,t-1}$ | 0.044*** | (0.006) | | | | |
| $N^S_{j,t-1}$ | 0.075*** | (0.017) | | | | |
| $P^S_{j,t-1}$ | −0.05 *** | (0.018) | | | | |
| $B^S_{j,t-1}$ | | | 0.025** | (0.011) | | |
| $Neg(B^S_{j,t-1})$ | | | −0.141*** | (0.018) | | |
| $B^{S*}_{j,t-1}$ | | | | | 0.020*** | (0.004) |
| $Neg(B^{S*}_{j,t-1})$ | | | | | −0.094*** | (0.007) |
| | Panel C: Returns $R_{j,t}$ | | | | | |
| $\sigma_{j,t-1}$ | −0.036*** | (0.010) | −0.036*** | (0.010) | −0.036*** | (0.010) |
| $R_{j,t-1}$ | −0.009*** | (0.003) | −0.009** | (0.003) | −0.009** | (0.003) |
| $R_{M,t-1}$ | −0.04 *** | (0.006) | −0.04 *** | (0.006) | −0.04 *** | (0.006) |
| $VIX_{t-1}$ | −0.000*** | (0.000) | −0.000*** | (0.000) | −0.000*** | (0.000) |
| $Ind_{j,t-1}$ | −0.000 | (0.000) | | | | |
| $N^S_{j,t-1}$ | −0.001 | (0.001) | | | | |
| $P^S_{j,t-1}$ | 0.001 | (0.001) | | | | |
| $B^S_{j,t-1}$ | | | 0.002*** | (0.001) | | |
| $Neg(B^S_{j,t-1})$ | | | −0.004*** | (0.001) | | |
| $B^{S*}_{j,t-1}$ | | | | | 0.001*** | (0.000) |
| $Neg(B^{S*}_{j,t-1})$ | | | | | −0.001*** | (0.000) |

*** refers to a $p$ value less than 0.01, ** refers to a $p$ value more than or equal to 0.01 and smaller than 0.05, and * refers to a $p$ value more than or equal to 0.05 and less than 0.1. Values in parentheses are standard errors.

**Table 11:** Time-Lagged Panel Regression Results on Sentence Level

TXTPanelTimeLagged.R

## References


Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, 71(2):579–625.

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

Bachelier, L. (2006). *Louis Bachelier's theory of speculation the origins of modern finance*. Princeton University Press, Princeton.

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

Bettman, J. L., Hallett, A. G., and Sault, S. (2011). Rumortrage. In *Finance and Corporate Governance Conference*.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python*. O'Reilly, Beijing; Cambridge.

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

Bondt, W. F. M. D. and Thaler, R. (1985). Does the stock market overreact? *Journal of Finance*, 40(3):793.

Brown, K. C., Harlow, W., and Tinic, S. M. (1988). Risk aversion, uncertain information, and market efficiency. *Journal of Financial Economics*, 22(2):355–385.

Chen, H., De, P., Hu, Y. J., and Hwang, B.-H. (2014). Wisdom of crowds. *Review of Financial Studies*, 27(5):1367–1403.

Chen, S. H. (2014). *Advances in computational social science: the fourth world congress*. Springer, New York.

Chen, Z., Daigler, R. T., and Parhizgari, A. M. (2006). Persistence of volatility in futures markets. *Journal of Futures Markets*, 26(6):571–594.

Das, S. R. and Chen, M. Y. (2007). Yahoo! for amazon. *Management Science*, 53(9):1375–1388.

Duckett, J. (2011). *HTML & CSS: design and build websites*. Wiley, Indianapolis.

Fama, E. F. (1970). Efficient capital markets. *Journal of Finance*, 25(2):383.

Garman, M. B. and Klass, M. J. (1980). On the estimation of security price volatilities from historical data. *Journal of Business*, pages 67–78.

Girard, E. and Biswas, R. (2007). Trading volume and market volatility. *Financial Review*, 42(3):429–459.

Groß-Klußmann, A. and Hautsch, N. (2011). When machines read the news. *Journal of Empirical Finance*, 18(2):321–340.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of ACM SIGKDD*, pages 168–177.

Jennings, F. and Yates, J. (2009). Scrapping over data. *Journal of Intellectual Property Law & Practice*, 4(2):120–129.

Kim, S.-H. and Kim, D. (2014). Investor sentiment from internet message postings and the predictability of stock returns. *Journal of Economic Behavior & Organization*, 107:708–729.

Kiss, T. and Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

Lane, J., editor (2012). *Foundation Website creation with HTML5, CSS3, and JavaScript*. Springer, New York.

Li, Q., Wang, T., Li, P., Liu, L., Gong, Q., and Chen, Y. (2014). The effect of news and public mood on stock movements. *Information Sciences*, 278:826–840.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? *The Journal of Finance*, 66(1):35–65.

MacIntyre, R. (1995). Penn treebank tokenization on arbitrary raw text.

Malkiel, B. G. (2003). The efficient market hypothesis and its critics. *Journal of Economic Perspectives*, 17(1):59–82.

Manning, C. D. (1999). *Foundations of statistical natural language processing*. MIT Press, Cambridge.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Nann, S., Krauss, J., and Schoder, D. (2013). Predictive analytics on public data-the case of stock markets. *ECIS*, pages 102–115.

Park, J., Konana, P., Gu, B., Kumar, A., and Raghunathan, R. (2013). Information valuation and confirmation bias in virtual communities. *Information Systems Research*, 24(4):1050–1067.

Sabherwal, S., Sarkar, S. K., and Zhang, Y. (2011). Do internet stock message boards influence trading? *Journal of Business Finance & Accounting*, 38(9-10):1209–1237.

Shu, J. and Zhang, J. E. (2006). Testing range estimators of historical volatility. *Journal of Futures Markets*, 26(3):297–313.

Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., and Deng, X. (2013). Exploiting topic based twitter sentiment for stock prediction. In *ACL (2)*, pages 24–29.

Sprenger, T. O., Tumasjan, A., Sandner, P. G., and Welpe, I. M. (2014). Tweets and trades: the information content of stock microblogs: Tweets and trades. *European Financial Management*, 20(5):926–957.

Tetlock, P. C. (2007). Giving content to investor sentiment. *Journal of Finance*, 62(3):1139–1168.

Truyens, M. and Van Eecke, P. (2014). Legal aspects of text mining. *Computer Law & Security Review*, 30(2):153–170.

Wang, G., Wang, T., Wang, B., Sambasivan, D., Zhang, Z., Zheng, H., and Zhao, B. Y. (2014). Crowds on wall street. *arXiv preprint arXiv:1406.1137*.

Webster, J. J. and Kit, C. (1992). Tokenization as the initial phase in NLP. *Proceedings of Conference on Computational Linguistics*, 4:1106–1110.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings HLTEMNLP*, pages 347–354.

Wisniewski, T. P. and Lambe, B. (2013). The role of media in the credit crunch. *Journal of Economic Behavior & Organization*, 85:163–175.

Yu, S., Rentzler, J., and Tandon, K. (2010). Reexamining the uncertain information hypothesis on the s&p 500 index and SPDRs. *Review of Quantitative Finance and Accounting*, 34(1):1–21.

Zarowin, P. (1990). Size, seasonality, and stock market overreaction. *Journal of Financial and Quantitative Analysis*, 25(1):113.

Zhang, J. L., Härdle, W. K., Chen, C. Y., and Bommes, E. (2015). Distillation of news flow into analysis of stock reactions.

Zhang, X., Fuehres, H., and Gloor, P. A. (2012). Predicting asset value through twitter buzz. In *Advances in Collective Intelligence 2011*, pages 23–34. Springer.

Zhang, Y. and Swanson, P. E. (2010). Are day traders bias free? *Journal of Economics and Finance*, 34(1):96–112.

## Declaration of Authorship

I hereby confirm that I have authored this Master's thesis independently and without use of others than the indicated sources. All passages which are literally or in general matter taken out of publications or other sources are marked as such.

Berlin, March 1 2015

Elisabeth Bommes