Bommes, Elisabeth; Chen, Cathy Yi-Hsuan; Härdle, Wolfgang Karl

Working Paper

Textual Sentiment and Sector specific reaction

IRTG 1792 Discussion Paper, No. 2018-043

Provided in Cooperation with:
Humboldt University Berlin, International Research Training Group 1792 "High Dimensional Nonstationary Time Series"

Suggested Citation: Bommes, Elisabeth; Chen, Cathy Yi-Hsuan; Härdle, Wolfgang Karl (2018) : Textual Sentiment and Sector specific reaction, IRTG 1792 Discussion Paper, No. 2018-043, Humboldt-Universität zu Berlin, International Research Training Group 1792 "High Dimensional Nonstationary Time Series", Berlin

This Version is available at:
https://hdl.handle.net/10419/230754

IRTG 1792 Discussion Paper 2018-043

Textual Sentiment and Sector specific reaction

Elisabeth Bommes *
Cathy Yi-Hsuan Chen *
Wolfgang Karl Härdle *

* Humbdolt-Universität zu Berlin, Germany

This research was supported by the Deutsche Forschungsgemeinschaft through the International Research Training Group 1792 "High Dimensional Nonstationary Time Series".

http://irtg1792.hu-berlin.de
ISSN 2568-5619

International Research Training Group 1792

QUANTITATIVE FINANCE

Textual Sentiment and Sector specic reaction

Journal:	*Quantitative Finance*
Manuscript ID	Draft
Manuscript Category:	Research Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Haerdle, Wolfgang ; Humboldt-Universitat zu Berlin, C.A.S.E. - Center for Applied Statistics and Economics
Keywords:	Investor Sentiment, Attention Analysis,, Text mining, polarity, investor sentiment
JEL Code:	C81, G14, G17, G14
Abstract:	News move markets and contains incremental information about stock reactions. Future trading volumes, volatility and returns are a ected by sentiments of texts and opinions expressed in articles. Earlier work of sentiment distillation of stock news suggests that risk prole reactions might di er across sectors. Conventional asset pricing theory recognizes the role of a sector and its risk uniqueness that di ers from market or rm specic risk. Our research assesses whether incorporating the sentiment distilled from sector specic news carries information about risk proles. Textual analytics applied to about 600K articles leads us with lexical projection and machine learning to classication of sentiment polarities. The texts are scraped from ocial NASDAQ web pages and with Natural Language Processing (NLP) techniques, such as tokenization, lemmatization, a sector specic sentiment is extracted using a lexical approach and a nancial phrase bank. Predicted sentence-level polarities are aggregated into a bullishness measure on a daily basis and fed into a panel regression analysis with sector indicators. Supervised learning with hinge or logistic loss and regularization yields good prediction results of polarity. Compared with standard lexical projections, the supervised learning approach yields superior predictions of sentiment, leading to highly sector specic sentiment reactions. The Consumer Staples, Health Care and Materials sectors show strong risk prole reactions to negative polarity.

Textual Sentiment and Sector specific reaction*

Elisabeth Bommes†

Cathy Yi-Hsuan Chen‡

Wolfgang Karl Härdle§

*The authors gratefully acknowledge financial support from the Deutsche Forschungsgemeinschaft through the International Research Training Group IRTG 1792 "High Dimensional Non Stationary Time Series" and the Collaborative Research Center CRC 649 "Economic Risk".

†Humboldt-Universität zu Berlin, C.A.S.E. - Center for Applied Statistics and Economics, Unter den Linden 6, 10099 Berlin, Germany. E-mail: elisabeth.bommes@hu-berlin.de

‡ Ladislaus von Bortkiewicz Chair of Statistics, C.A.S.E. - Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

§Humboldt-Universität zu Berlin, C.A.S.E. - Center for Applied Statistics and Economics, Unter den Linden 6, 10099 Berlin, Germany. Visiting Professor in Sim Kee Boon Institute for Financial Economics, Singapore Management University, 90 Stamford Road, 6th Level, School of Economics, Singapore 178903.

Abstract

News move markets and contains incremental information about stock reactions. Future trading volumes, volatility and returns are affected by sentiments of texts and opinions expressed in articles. Earlier work of sentiment distillation of stock news suggests that risk profile reactions might differ across sectors. Conventional asset pricing theory recognizes the role of a sector and its risk uniqueness that differs from market or firm specific risk.

Our research assesses whether incorporating the sentiment distilled from sector specific news carries information about risk profiles. Textual analytics applied to about 600K articles leads us with lexical projection and machine learning to classification of sentiment polarities. The texts are scraped from official NASDAQ web pages and with Natural Language Processing (NLP) techniques, such as tokenization, lemmatization, a sector specific sentiment is extracted using a lexical approach and a financial phrase bank. Predicted sentence-level polarities are aggregated into a bullishness measure on a daily basis and fed into a panel regression analysis with sector indicators. Supervised learning with hinge or logistic loss and regularization yields good prediction results of polarity. Compared with standard lexical projections, the supervised learning approach yields superior predictions of sentiment, leading to highly sector specific sentiment reactions. The Consumer Staples, Health Care and Materials sectors show strong risk profile reactions to negative polarity.

Keywords: Investor Sentiment, Attention Analysis, Sector-specific Reactions, Volatility, Text Mining, Polarity

JEL Classifications: C81, G14, G17

1 Introduction

News are undoubtedly driving financial markets. In digital form news feeds are nowadays ubiquitous and massively available on a plethora of platforms in a wide spectrum of granularity scales. The size of this information pool makes it virtually impossible to process all the news relevant to certain financial assets since one runs automatically in a "noise" vs. "signal" conflict. Exceptions are of course scheduled events like central bank announcements for which many empirical studies on news impact are available. An early study is Rosa and Verga (2007), followed by Al-Rjoub (2016). All these approaches have limitations though since they concentrate on identifiable indicators (events like quarterly reports) or use specific automated linguistic algorithms.

Recent studies have looked at continuous news from an automated sentiment machine learning point of view, see Zhang et al. (2016), Cao et al. (2001), Das and Chen (2007), Chen et al. (2014), Rönnqvist and Sarlin (2017), Schumaker et al. (2012) and Guo et al. (2017). In summary the distilled sentiments have been discovered to be relevant to high frequency return, volatility and trading volume. Whereas it is shown in these papers that small investors' opinions contribute to stock markets and create in general "news-driven" stock reactions deeper sentence based analysis and a view on industry sectors are missing though. Indeed, a discovery in Zhang et al. (2016) on a possible sector specific behavior was that the health care and the finance sector displayed quite different impulses. This observation and the advance in Natural Language Processing (NLP) is in fact the motivation of the research carried out here. More specifically we are interested in the following questions:

1. Are there sector specific reactions on volatility, returns?
2. Is there an information gain by looking at sentence based news?
3. Are these reactions on an intra day or lagged time level identical?

In order to answer these we not only carry out the standard projection techniques of the dominantly used sentiment lexica: the BL by Hu and Liu (2004) and LM by Loughran and McDonald (2011) lexica. but also apply newer p-gram sentence based techniques. We rely on exploiting these different projections, and use an extended data set that in the meanwhile amounts to more than 580K NASDAQ articles from 2012.1 to 2016.12.

In previous work, either non-public training sets were created as or sentiment classification in a financial setting as e.g. in Antweiler and Frank (2004) or unsupervised sentiment is projected with lexica. We close this gap by comparing sentiment predicted by a supervised approach labeled as SM and based on the Gold Standard Corpus (GSC) for financial sentiment by Malo et al. (2014) with the de facto most common lexical approach. Supervised modeling of sentiment overcomes several of the limitations of lexical projections: First and foremost, positive and negative words are not necessarily weighted equally. We also lemmatize the individual token as a more sophisticated and interpretable alternative to stemming and incorporate p-grams in addition to the widely used 1-gram bag of words approach.

While employing stratified 5-fold cross-validation to avoid overfitting, we estimate and evaluate more than 65,000 candidate models to determine the best combination of model tuning parameters such as the loss function and regularization norm. Comparability of the sentence-level supervised sentiment with lexical sentiment is ensured by mapping the lexical approach to sentence polarity. For the further panel analysis, sentences are aggregated on a per day and company basis by a single bullishness measure introduced by Antweiler and Frank (2004) as well as the fractions of positive and negative sentiment.

Our main findings are as follows. Supervised sentence based polarity calculation outperforms lexical approaches in terms of accuracy, precision and recall on the manually labeled training data. Furthermore, the estimated parameters of the SM are consistently significant in the context of contemporaneous regression models in contrast to those of the standard lexical based extractions. We find indeed, using advanced machine learning techniques, significant differences between sector specific reactions. As an example, the effect of negative sentiment on volatility is significantly larger for the Health Care sector than for Financials while the opposite applies for returns. We also conclude that the bullishness variable, which combines positive and negative sentiment in a single measure, has significant disadvantages compared with fractions as explanatory variable. This result holds when introducing the negative part of the bullishness as an additional variable to account for asymmetric effects.

The algorithms have been programmed in Python and R and the natural language processing was carried out with the Python module "Natural Language Processing Toolkit" by Bird et al. (2009). The algorithms are available as quantlets on quantlet.de and the data

set is available for research purposes at the Humboldt Lab for Empirical and Quantitative Research at Humboldt Universität zu Berlin, Germany.

The next Section 2 presents the data in detail and gives the exact steps for calculation of the polarities, resulting finally in the fractions for positive and negative sentiment as well as a measure for bullishness. Section 3 enters into the comparison of the lexicon based projections vs. the supervised learning techniques based sentences level. Section refse:panel presents the panel regression results to compare the different sentiment calculation approaches. Section 5 concludes.

2 Data

We consider news articles that are available at the Nasdaq news platform from 6 Jan 2009 to 29 December 2016. The textual data from this source was acquired by Zhang et al. (2016) via an automatic web scraper and is available for academic purposes at the Research Data Center of the Collaborative Research Center 649 at Humboldt-Universität zu Berlin. While the data origin suggests that only companies traded on the Nasdaq exchange are discussed, also articles about companies listed at other exchanges are available.

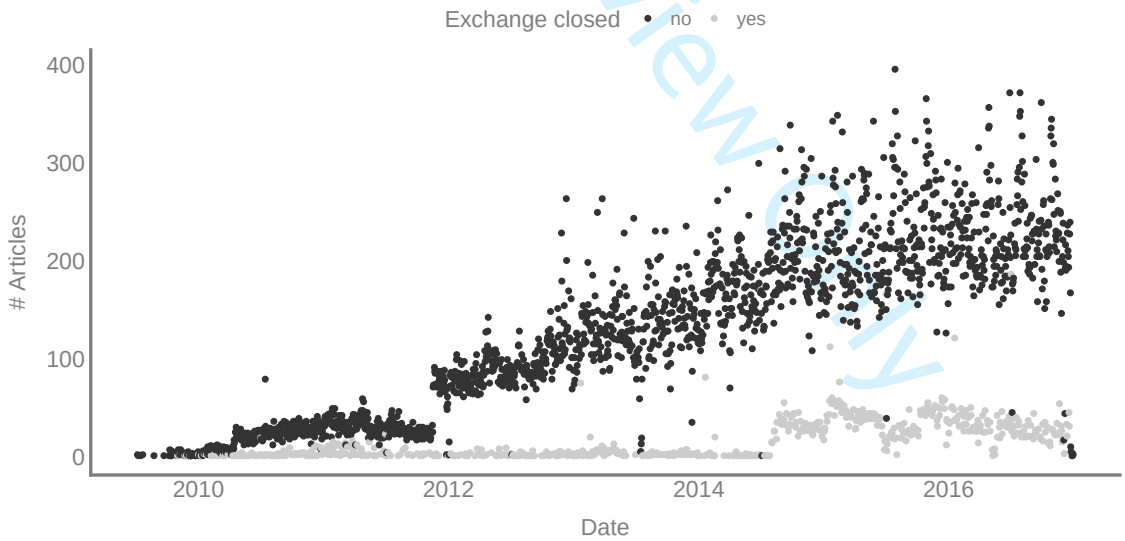


Figure 1: Number of Articles about S&P 500 listed Companies per Day

In total, there are 581,709 articles during the discussed time frame. suggests a tremendous increase in the number of articles as Zhang et al. (2016) only discuss 116,691 articles as of October 2014. However, a good portion of the collected articles either discuss stocks that are not listed in the S&P 500 or they relate to e.g. currencies and commodities.

Restricting the data set to articles about at least one company listed in the S&P 500 index results in 239,381 articles.

Figure 1 illustrates the number of published articles per day over time. One can observe a structural break between 2011 and 2012 such that the average number of articles per day in 2012 is more than twice as much as in 2011. This may be due to an establishment phase of the news platform but specific reasons remain unknown to us. We limit our further analysis to observations after 2011 to avoid this structural break interfering with the model. Furthermore, this limitation also has the benefit that the publishing rate is substantially higher than in the earlier periods. Our initial concerns regarding days with a very low publishing rate are unjustified as these days usually coincide with weekends or public holidays and thus, trading does not take place.

As shown in Zhang et al. (2016), the attention by the media differs among companies and it is proposed to investigate the attention ratio as proxy of media interest. They define the attention ratio as

$$AR_i = T^{-1} \sum_{t=1}^{T} \mathbf{I}\left(c_{i,t} > 0\right) \tag{1}$$

with $c_{i,t}$ being the number of articles for company i on day t and the total number of days in the data set T. Thus, this ratio represents the proportion of days on which a company has at least one written article about them. Table 1 shows that 80% of the S&P 500 companies do appear in the investigated media subsample less than every second day. As Zhang et al. (2016) concluded that the attention ratio has an impact on the significance of parameters in a panel regression setting, we limit our further analysis to the 100 S&P 500 companies with the highest attention ratios in order to eliminate this effect.

Quantile	0%	20%	40%	60%	80%	100%
Attention Ratio	0.01	0.18	0.22	0.30	0.44	0.99

Table 1: Quantiles of Attention Ratio for all S&P 500 companies

In the following, we take a closer look at available data regarding the individual industry sectors. Here, we determine each company's industry based on the Global Industry Classification Standard (GICS) sector code. Table 12 in the Appendix gives an overview about this industry taxonomy.

Sector	Attention Ratio Min	Q1	Q2	Q3	Max	Companies
Consumer Discretionary	0.448	0.523	0.630	0.737	0.929	19
Consumer Staples	0.443	0.500	0.521	0.622	0.871	10
Energy	0.448	0.512	0.534	0.697	0.854	8
Financials	0.464	0.616	0.686	0.891	0.979	13
Health Care	0.443	0.512	0.583	0.636	0.841	13
Industrials	0.458	0.522	0.577	0.661	0.857	13
Information Technology	0.444	0.528	0.655	0.848	0.991	18
Materials	0.533	0.585	0.637	0.640	0.643	3
Telecommunication Services	0.871	0.885	0.899	0.913	0.927	2
Utilities	0.463	0.463	0.463	0.463	0.463	1

Table 2: Attention Ratio of 100 Companies by Sector. Q1, Q2 and Q3 represent 25%, 50% and 75% quantile, respectively.

Removing the companies with a low attention ratio leads to the distribution of the AR across sectors shown in Table 2. It is obvious that Utilities as a sector does not receive a lot of media coverage as there is only one company representing this sector. There are also only two companies with Telecommunication Services as sector which may be due to the fact, that there is only a total of thee such companies represented in the S&P 500. Both, the Utilities and Telecommunication Services sectors are excluded from the data sample. The AR quantiles of the remaining sectors are quite comparable with the exception that companies in Financials, Information Technology and Consumer Discretionary get more media coverage than in the remaining sectors while Materials get slightly less.

Table 3 goes into more detail regarding the distribution of articles across sectors and the company with the highest AR in each sector. While one observes by comparing Table 3 with the previous Table 2 that Information Technology and Consumer Discretionary have about the same number of companies, Information Technology has about 50% more news items and thus the media interest in this sector seems to be higher. The distribution over time is roughly the same for each sector. Financials and Materials get a higher coverage in the early data set while Consumer Discretionary, Energy and Health Care receive more articles in late 2014 and 2015.

	Number (in k)			Date			Articles per Day					Sentences per Article				
Sector, Stock Symbol	Articles	Sentences	Words	Q1	Q2	Q3	$\widehat{\mu}$	$\widehat{\sigma}$	Q1	Q2	Q3	$\widehat{\mu}$	$\widehat{\sigma}$	Q1	Q2	Q3
Consumer Discretionary	30.36	991.89	19,492.67	2013-07-11	2014-08-22	2015-06-03	29.59	13.94	19	27	38	32.68	25.77	18	26	43
Consumer Staples	12.21	366.71	7,374.95	2013-06-20	2014-08-01	2015-04-17	11.96	6.56	7	11	16	30.03	22.58	15	24	41
Energy	10.41	300.14	6,231.60	2013-05-07	2014-08-26	2015-05-27	10.25	5.78	6	9	13	28.83	21.96	14	24	40
Financials	34.57	769.21	14,776.58	2013-05-29	2014-05-21	2015-03-27	33.73	24.05	20	28	40	22.25	18.57	10	15	28
Health Care	16.95	480.34	9,520.63	2013-08-21	2014-08-27	2015-05-11	16.60	9.28	9	15	23	28.35	23.64	15	22	36
Industrials	16.44	450.48	8,840.50	2013-06-18	2014-07-17	2015-05-20	16.11	8.24	10	15	21	27.39	20.37	15	22	36
Information Technology	44.12	1,447.70	28,094.29	2013-06-13	2014-07-30	2015-04-28	43.00	18.04	30	41	54	32.82	23.91	19	28	43
Materials	3.82	100.73	1,966.75	2013-04-08	2014-04-14	2015-04-13	4.16	3.11	2	3	5	26.39	21.74	13	20	34
Telecommunication Services	5.88	164.05	3,264.24	2013-05-16	2014-07-17	2015-05-13	5.92	3.60	3	5	8	27.92	19.94	16	23	36
Utilities	0.78	17.59	332.45	2013-05-29	2014-08-09	2015-05-11	1.63	1.00	1	1	2	22.66	17.95	11	15	29
AMZN	4.75	174.25	3,452.61	2013-10-17	2014-11-06	2015-07-31	4.96	3.72	2	4	7	36.68	27.25	21	31	48
WMT	2.80	89.85	1,851.18	2013-05-08	2014-06-27	2015-04-15	3.12	2.38	2	3	4	32.11	23.05	18	27	43
XOM	2.46	70.74	1,486.19	2013-06-20	2014-08-29	2015-05-19	2.80	1.82	1	2	4	28.72	23.00	14	23	40
JPM	6.19	133.76	2,600.66	2013-06-18	2014-06-10	2015-04-15	6.24	4.83	3	5	8	21.61	18.76	10	15	26
GILD	2.34	68.09	1,316.48	2014-05-01	2014-09-30	2015-06-18	3.58	2.86	1	3	5	29.05	22.45	17	22	37
BA	2.43	76.44	1,506.44	2013-06-17	2014-06-20	2015-04-24	2.75	1.92	1	2	4	31.47	23.43	17	24	41
AAPL	11.85	429.58	8,376.21	2013-07-08	2014-10-03	2015-05-18	11.61	7.43	6	10	15	36.25	25.50	22	32	46
DD	1.33	31.83	615.18	2013-03-11	2014-04-22	2015-05-04	2.00	1.48	1	2	2	24.00	21.82	12	17	29
T	3.23	92.75	1,824.22	2013-05-06	2014-06-19	2015-05-26	3.38	2.10	2	3	4	28.70	21.06	16	24	37
DUK	0.78	17.59	332.45	2013-05-29	2014-08-09	2015-05-11	1.63	1.00	1	1	2	22.66	17.95	11	15	29

Q1, Q2 and Q3 represent 25%, 50% and 75% quantile, respectively. $\widehat{\mu}$ and $\widehat{\sigma}$ denote the empirical average and standard deviation.

Table 3: Summary Statistics of News Articles about 100 Companies

There is no significant difference of sentences per article across either sectors or companies which is mainly due to the fact, that the standard deviation of sentences is quite high. All in all, we can conclude that differences among sectors in their reaction to sentiment in the articles may not be due to distributional differences in e.g. the volume of articles over time or the length of the mentioned articles.

Stock specific data such as the S&P 500 constituents and daily prices is collected from Bloomberg and Compustat. Compustat is used to gather Global Industry Classification Standard (GICS) sector for these assets. In the following, two stock reactions are considered: volatility 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 may be improved by using high-low data and define the range-based measure of volatility for company i on day t as

$$\begin{aligned} \sigma_{i,t} &= 0.511(u-d)^2 - 0.019\left\{c(u+d) - 2ud\right\} - 0.383c^2 \\ \text{with } u &= \log(P_{i,t}^H) - \log(P_{i,t}^O), \\ d &= \log(P_{i,t}^L) - \log(P_{i,t}^O), \\ c &= \log(P_{i,t}^C) - \log(P_{i,t}^O), \end{aligned} \tag{2}$$

with $P_{i,t}^H$, $P_{i,t}^L$, $P_{i,t}^O$, $P_{i,t}^C$ being 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 on daily level. Subsequently, the Garman and Klass range-based measure of volatility is used in the further analysis.

Furthermore, the returns are calculated as

$$r_{i,t} = \log(P_{i,t}^C) - \log(P_{i,t-1}^C) \tag{3}$$

for company i on day t.

3 Sentiment

3.1 Unsupervised Projection

In recent years, several lexica have been assembled for the purpose of sentiment projection such as Hu and Liu (2004) and Loughran and McDonald (2011) referred to as BL and LM, respectively. While a word-level approach to determine investor sentiment is common in Economics, it is suggested that sentiment analysis on sentence or even phrase level is a superior regarding the projection accuracy by e.g. Wiebe and Riloff (2005) and Wilson (2005).

Let $D = (d_1, d_2, \ldots, d_n)_{n\in\mathbb{N}}$ be a finite sequence of words d and the corresponding set $L = \{D\} = \{l_1, l_2, \ldots, l_m\}$ with $|L| = m$ as a lexicon containing all words that appear in D. The further steps are tailored for D being a single sentence but can be easily adjusted for a word or document based sentiment projection. The number of appearances of each distinct word l_i $(i = 1, \ldots, m)$ can be counted by

$$c_i = c(D, l_i) = \sum_{j=1}^{n} \mathbf{I}\left(d_j = l_i\right) \tag{4}$$

as the order of each word in D does not change the count values.

Furthermore, let L_o be a polarity lexicon with $o \in \{pos, neut, neg\}$ corresponding to positive, neutral and negative words, respectively. Note that $L_{neut} = (L_{pos} \cup L_{neg})^{\mathsf{c}}$ as it holds for any suitable selection of lexica such that $L_i \cap L_j = \emptyset$ for $i \neq j$ and $i, j \in o$. The count values of polarity words are thus calculated by

$$w_o = w(D, L_o) = \sum_{i=1}^{m} \mathbf{I}\left(l_i \in L_o\right) c_i \ . \tag{5}$$

Formula 5 only considers polarity on a word level. This approach may be oversimplified as Polanyi and Zaenen (2006) state the importance of valence shifters for the interpretation of a text. Valence shifters are words that modify another word such that its connotation is flipped. Quite an obvious example are negation words, as "not good" has clearly not a positive meaning.

In practice, negation is often handled by looking at the n-gram, a sequence of n words, around $d \in L_o$. One can see that the position in the text matters for such an approach and words may not be re-ordered. Thus, if the distance between a sentiment word and a negation word is less than a pre-specified threshold, the polarity of the word is inverted as suggested in e.g. Hu and Liu (2004). Formally, if we specify the number of words to consider that appear before and after a polarity word as k, we can calculate the number of shifted polarity words as

$$v_o = v(D, L_o, L_v) = \sum_{i=1}^{n} \mathbf{I}\left(d_i \in L_o\right) \mathbf{I}\left(s_i \in \mathbb{O}\right) \tag{6}$$

with $s_i = \sum_{j=\max(i-k,1)}^{\min(i+k,n)} \mathbf{I}\left(d_j \in L_v\right)$ and $\mathbb{O} = \{2h+1 : h \in \mathbb{N}\}$. As a convention, words are only shifted when there is an odd number of valence shifters around the polarity word. As for a suitably small value of k, the case $k \leq 1$ occurs predominantly. In the following, we consider $k = 3$ and fix $L_v = \{$ `"n't", "not", "never", "no", "neither", "nor", "none"` $\}$ as shifting lexicon.

Up to now, we established how to count the number of (shifted) polarity words in a sentence. The polarity of the whole sentence can now simply given by

$$S_L = S(D, L_o, L_v) = sgn(w_{pos} + v_{neg} - w_{neg} - v_{pos}) \tag{7}$$

with $S_L \in \{1, 0, -1\}$ which corresponds to $\{pos, neut, neg\}$ in terms of sentiment.

3.2 Supervised Projection

Despite being widely applied, polarity lexica have the common downfall, that they only take sentiment on a word-level into account while neglecting the overall phrase structure. Malo et al. (2014) provide the financial phrase bank, a data set containing roughly 5,000 financial sentences which have been manually labeled by 16 annotators as positive, neutral and negative. Three typical sentences from the phrase bank are given in Example 1 with the tagged polarity indicated by a preposed `@` after the sentence.

```
1 With the new production plant the company would increase its capacity
    to meet the expected increase in demand and would improve the use
    of raw materials and therefore increase the production
    profitability. @positive
```

2 According to Gran, the company has no plans to move all production to Russia, although that is where the company is growing. @neutral
3 The company slipped to an operating loss of EUR 2.6 million from a profit of EUR 1.3 million. @negative

Example 1: Polarity Sentences from Financial Phrase Bank

Both, the sentences from the phrase bank and the Nasdaq text corpus are pre-processed as follows by by using the Python Natural Language Toolkit (NLTK) by Bird et al. (2009). While some of the steps are not necessary for the phrase bank and e.g. sentence boundary detection as the end of each sentence is clearly identifiable by a new line, they are needed to bring the Nasdaq corpus into the same form as the phrase bank. Jurafsky and Martin (2009) lists the typical parts of this text normalization in three steps as the (1) segmentation of words, (2) the normalization of word formats and (3) the segmentation of sentences.

Steps (1) and (3) are executed by word- and sentence-level tokenization, respectively. Word tokenization is the process of breaking text down into its word based units and according to Webster and Kit (1992) the automatic analysis of text is impossible without it. A simplistic approach of using space delimiters as token identifiers is not sufficient as e.g. `company's` consists of the word `company` and the Anglo-Saxon genitive of nouns. Thus, we apply the more sophisticated Penn Treebank tokenizer by MacIntyre (1995) which is built on regular expressions and able to handle the mentioned case as well as common contractions of words. Similarly, sentence boundaries should not be detected by identifying punctuation such as `"."`, `"?"` and `"!"` due to e.g. abbreviations, initials and specific cases of numbers. The Punkt tokenizer by Kiss and Strunk (2006) is an unsupervised approach to detect these sentence boundaries in a text and a pre-trained version is included in NLTK.

Furthermore, each sentence is converted to lower case, non alphabetic characters are removed and each word is lemmatized to account for step (2). In contrast to stemmers like the one by Porter (1980), lemmatization takes the word's part of speech into account while converting the inflected word into its root. As a simple example, using the Porter stemmer on a word such as `was` would not change it while lemmatizing it leads to `be` as outcome.

1 The profit of Apple increased.

```
2 The profit of the company decreased.
```

Example 2: Simple Sentences

For instance, look at the sentences in Example 2. Here, the normalization results in `("the", "profit", "of", "apple", "increase")` as well as `("the", "profit", "of", "the", "company", "decrease")`.

After this natural language processing, the sentences still have to be brought in a format with count values such that statistical modeling is possible. In the following, we use a data processing pipeline which is built with scikit-learn, a Python machine learning library by Pedregosa et al. (2011).

Next, a count vectorizer is employed to bring the lemmatized tokens of all sentences in a numerical matrix representation. Here, a lot of different options regarding the n-gram range, the removal of stop words, as well as minimal and maximal document frequency need to be considered. In the following, we explain these tuning parameters on the basis of sentence 1 in Example 2 and the range in which we tune them.

Some words of the English language such as `"the"` appear so often that they hardly bear any meaning. Stop word removal aims to reduce noise by identifying and eliminating such words based on a fixed word list. The Glasgow Information Retrieval Group compiled such a list which is available for download on their web page as well as part of NLTK. The removal of the mentioned stop words would lead to `("profit", "apple", "increase")`, a more dense representation of the sentence's meaning. One of the drawbacks of such an approach can be that several polarity shifters are part of this collection.

On the other hand, words with little meaning may be domain specific instead of universal, commonly referred to as corpus stop words. Hence, another approach is to remove words that appear either too frequently or very infrequent in the given collection of text. Here, we consider the maximum and minimum thresholds of $TF_{max} = \{0.85, 0.90, 0.95, 1.0\}$ and $TF_{min} = \{0.00, 0.01, 0.05, 0.1\}$ respectively, while building the vocabulary.

In addition to building the vocabulary from single tokens, also known as 1-grams, we consider the option of taking 2-grams into account. To refer to our example, the 2-grams would be `("the profit", "profit of", "of apple", "apple increase")` without stop word removal and `("profit apple", "apple increase")` with prior filtering of (corpus) stop

words.

Let us assume for now, that we apply a count vectorizer with 1- and 2-grams as well as stop word removal. Thus, the set of documents, here sentences, is $D = \{d_j\}_{j=1}^{n} = \{d_1, d_2\}$ and our vocabulary are the terms $T = \{t_i\}_{i=1}^{m} = \{$ `"profit", "apple", "company", "increase", "decrease", "profit apple", "apple increase", "profit company", "company decrease"`$\}$. After this, the term-document-matrix is calculated, resulting in

$$W^{\top} = \begin{array}{c} \\ \text{profit} \\ \text{apple} \\ \text{company} \\ \text{increase} \\ \text{decrease} \\ \text{profit apple} \\ \text{apple increase} \\ \text{profit company} \\ \text{company decrease} \end{array} \begin{array}{c} \begin{array}{cc} d_1 & d_2 \end{array} \\ \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \end{array} \tag{8}$$

with the term frequencies $tf_{i,j}$ as the raw term frequencies. As we can observe, these term frequencies could be used as term weights for further modeling as e.g. seen in Luhn (1957) such that words that appear more often reflect a higher meaning in the text. However, Sparck Jones (1972) figure that terms that appear evenly throughout all documents, here sentences, are not as helpful for discrimination purposes as terms that appear in only a couple of documents. Thus, we normalize the matrix by re-weighting the term-frequency with the inverse document-frequency (tf-idf) which is given by

$$tf\text{-}idf(i,j) = tf(i,j) \;\; idf(i,j) \tag{9}$$

and

$$idf(i,j) = \log[(1+n)\{1 + \sum_{i=1}^{n} \mathbf{I}\left(tf_{i,j} > 0\right)\}^{-1}] + 1 \tag{10}$$

resulting in

$$X = \begin{pmatrix} tf\text{-}idf(1,1) & tf\text{-}idf(1,2) & tf\text{-}idf(1,3) & \dots & tf\text{-}idf(1,m) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ tf\text{-}idf(n,1) & tf\text{-}idf(n,2) & tf\text{-}idf(n,3) & \dots & tf\text{-}idf(n,m) \end{pmatrix} \tag{11}$$

Note that Equation 10 might differ from the standard textbook notation due to the added constant in the nominator and denominator to prevent zero divisions.

As for the modeling part, note that we face a multi-class classification problem as the manually classified sentiment by Malo et al. (2014) is given for each sentence as $y_j \in \{-1, 0, 1\} = \{negative, neutral, positive\}$. In a one-vs-all classification scheme (OVA), we can simply estimate three individual binary classifiers to discriminate between one class and the group the remaining classes. As an example, y_j with j being positive sentiment as target class would be equal 1 if the classified sentiment is indeed positive and -1 otherwise. On the other hand, it is also possible to discriminate between the classes in an all-vs-all approach such that we would estimate $\binom{3}{2}$ binary classifiers in our given problem. As stated by Rifkin and Klautau (2004), the all-vs-all classification scheme has not substantial advantages over the simpler one-vs-all scheme as long as the binary classifiers are regularized in a sensible way. Hence, we focus on the OVA approach with regularized linear models (RLM) and implement it with scikit-learn.

The linear scoring function $s(X) = \beta^\top X$ with $\beta \in \mathbb{R}^m$is now calibrated with the regularized training error

$$n^{-1} \sum_{i=1}^{n} L(y_i, s(X_i)) + \lambda R(\beta) \tag{12}$$

with $L(\cdot)$ as loss function, $R(\cdot)$ as regularization term and hyperparameter $\lambda \geq 0$. Two candidates for the loss function are the Hinge loss, leading to a linear support vector machine model (SVM) as well as the well known logistic loss with the Hinge loss given by

$$L_{Hinge}(y_i, s(X_i)) = \max(0,\, 1 - s(X_i)y_i). \tag{13}$$

Furthermore, we also consider the squared Hinge and perceptron loss functions. As for the regularization, we consider L_1 with $R(\beta, L_1) = \sum_{i=1}^{p} |\beta_i|$ and L_2 regularization with $R(\beta, L_2) = p^{-1} \sum_{i=1}^{p} \beta_i^2$ as well as an elastic net, the combination of both as described by

Zou and Hastie (2005), and no regularization at all.

We employ stratified 5-fold cross validation to avoid overfitting. Furthermore, we oversample sentences with positive and negative sentiment in the Malo training set to obtain a balanced sample and control for the trade off between the type 1 and type 2 error as described by Härdle et al. (2009). Additional care is taken into account for the oversampling in combination with the 5-fold crossvalidation such that the re-sampling of a sentence stays in the same fold as the original sentence. Using this approach we estimate more than 65,000 candidate models and select the best supervised model S_M in terms of mean accuracy across folds. The resulting model has the following specifications

- Count vectorizer: $TF_{min} = 0$, $TF_{max} = 0.85$, no stop words removal and the inclusion of 1-grams and 2-grams
- tf-idf(1,1): L_2 norm
- RLM: $\hat{\lambda} = 0.0001$, L_{Hinge}, $R(\beta, L_1)$

The confusion matrices of both the upsampled version and the data with the original sentiment distribution can be found in Table 4 and the accuracies are 0.8 and 0.82, respectively.

Pred / True	S_M				S_M with Oversampling			
	-1	0	1	Total	-1	0	1	Total
-1	**389**	67	58	514	**1,983**	289	254	2,535
0	96	**2,134**	305	2,535	96	**2,134**	305	2,535
1	54	198	**916**	1,168	105	469	**1,961**	2,535
Total	539	2,399	1,279	4,217	2,184	2,901	2,520	7,605
Precision	0.72	0.89	0.72		0.78	0.73	0.78	
Recall	0.76	0.84	0.78		0.91	0.84	0.77	

Table 4: Confusion Matrix of Supervised Model

Furthermore, we also project the lexical sentiment onto the sentences such that we can employ a more detailed comparison between our unsupervised and supervised methods. The overall accuracies of S_{BL} and S_{LM} are 0.58 and 0.63, respectively, resulting in classifiers which are better than a random classification in the categories.

The further results are given in the confusion matrices in Table 5 and one can observe that

True \ Pred	S_{BL}			S_{LM}			
	-1	0	1	-1	0	1	Total
-1	**214**	268	32	**213**	289	12	514
0	203	**1,786**	546	200	**2,187**	148	2,535
1	89	627	**452**	111	772	**285**	1,168
Total	506	2,681	1,030	524	3,248	445	4,217
Precision	0.42	0.67	0.44	0.41	0.67	0.64	
Recall	0.42	0.71	0.39	0.41	0.86	0.24	

Table 5: Confusion Matrix of Lexical Projection

the supervised projection seems to have substantial advantages over the lexical projection regarding overall accuracy, precision and recall. Only S_{LM} seems to have a slight edge over S_M regarding the recall of the neutral sentiment classifier but however, its precision is much lower.

3.3 Sentiment Measures

In contrary to the projection on a sentence-level in Sections 3.1 and 3.2, sentiment measures aim to summarize the polarity of multiple sentences. Let $S = (S_1, S_2, \ldots, S_n)_{n\in\mathbb{N}}$ be the projected sentiment from either an unsupervised or supervised method for a n sentences.

One way to measure the sentiment of a document is by using the fractions of polarity which is e.g. used by Chen et al. (2014) and Zhang et al. (2016) on a word-level by calculation the percentage of words that are either positive or negative. Transferred to sentence-level sentiment, the fractions of polarity are then given by

$$PF = n^{-1}\sum_{i=1}^{n}\mathbf{I}\left(S_i = 1\right) \quad \text{and} \quad NF = n^{-1}\sum_{i=1}^{n}\mathbf{I}\left(S_i = -1\right). \tag{14}$$

Antweiler and Frank (2004) go one step further and combine both, negative and positive sentiment into one measure of bullishness which may be defined by means of the fractions in Equation 14 and is given by

$$B_A \;=\; \log(1+PF) \;-\; \log(1+NF). \tag{15}$$

One can easily observe, that $B_A < 0$ holds if the polarity of the text is negative while

$B_A = 0$ indicates neutrality and $B_A > 0$ suggests a positive polarity. Furthermore, as $B_A \in [\log(0.5), \log(2)]$ holds due to $PF, NF \in [0, 1]$, we can scale the bullishness by

$$B = \log(2)^{-1} B_A, \tag{16}$$

such that $B \in [-1, 1]$ holds which simplifies the interpretation of the calculated values. Due to the asymmetric reaction of stock indicators discussed in Zhang et al. (2016), we also specify the negative part of B by defining

$$BN = \mathbf{I}\big(B < 0\big)\, B. \tag{17}$$

In the following, we calculate $PF_{i,t}$, $NF_{i,t}$, $B_{i,t}$ and $BN_{i,t}$ for the BL and LM lexica as well as the supervised projection SM with i referring to the company and t being the date. This is done for the 100 companies selected in Section 2. One can already get a clue in Table 6 that the bullishness B is not distributed symmetrically around zero as the the estimated medians are clearly positive. In comparison, the statistics regarding the stocks' returns indicate that their distribution is fairly symmetrical around zero.

Variable		Min	Q1	Q2	Q3	Max	$\widehat{\mu}$	$\widehat{\sigma}$
B	BL	−1.00	0.00	0.09	0.26	1.00	0.14	0.19
	LM	−1.00	0.00	0.00	0.12	1.00	0.04	0.16
	SM	−0.87	0.00	0.15	0.30	1.00	0.17	0.18
NF	BL	0.00	0.00	0.08	0.18	1.00	0.10	0.11
	LM	0.00	0.00	0.07	0.16	1.00	0.10	0.11
	SM	0.00	0.00	0.03	0.09	0.83	0.05	0.07
PF	BL	0.00	0.00	0.26	0.38	1.00	0.23	0.20
	LM	0.00	0.00	0.13	0.21	1.00	0.13	0.12
	SM	0.00	0.00	0.20	0.31	1.00	0.19	0.17
Returns		−0.41	−0.01	0.00	0.01	0.38	0.00	0.02

Q1, Q2 and Q3 represent 25%, 50% and 75% quantile, respectively. $\widehat{\mu}$ and $\widehat{\sigma}$ denote the empirical average and standard deviation.

Table 6: Summary Statistics of Sentiment Measures

Thus, the news articles may contain a positive bias, either to avoid libel lawsuits, to

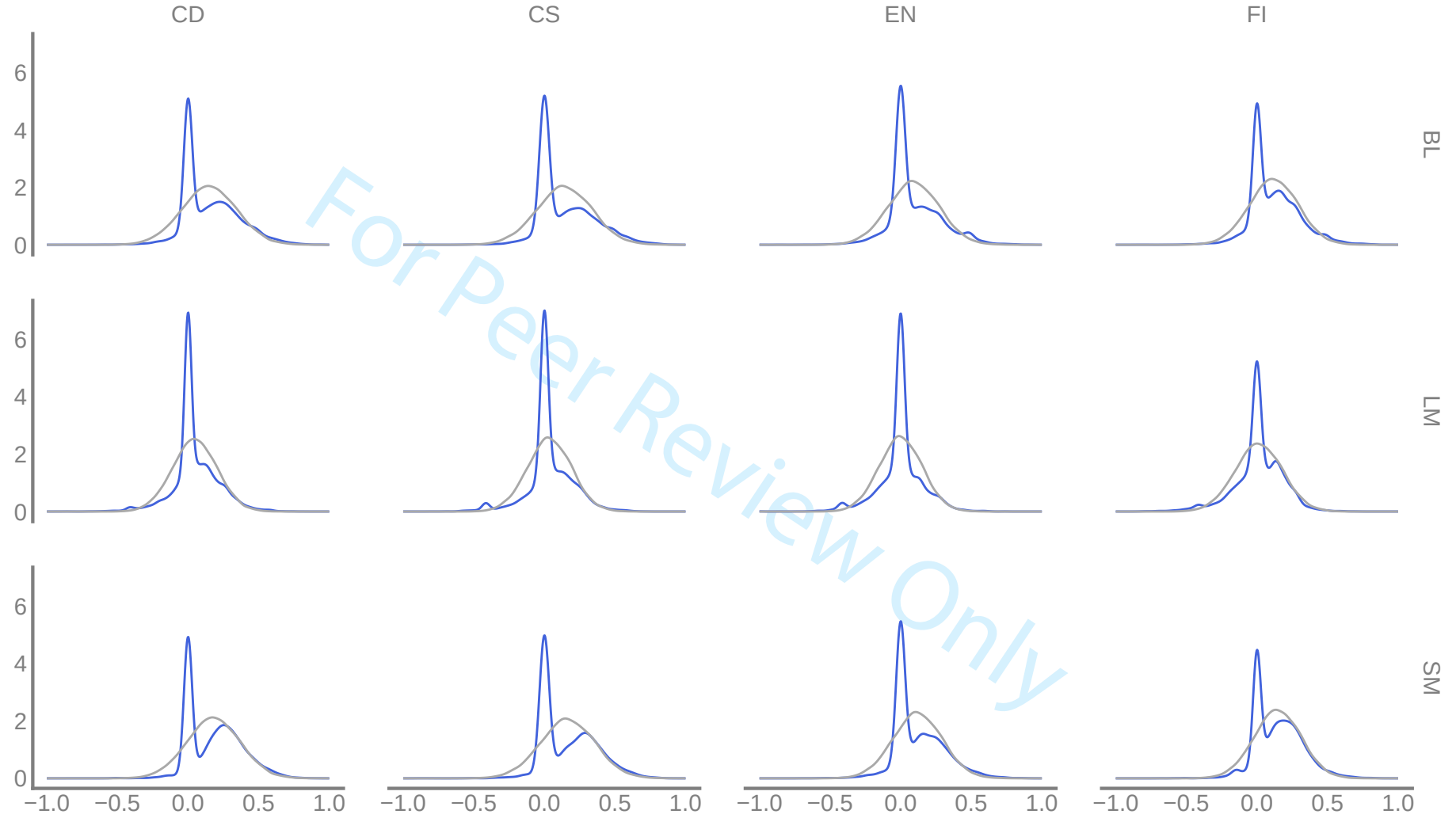
motivate investing in certain stocks or due to biased opinions of the authors. One recent example of a company filing a lawsuit against a news outlet is Murray Energy Corporation accusing the New York times of defamation. On the other hand, the U.S. Security and Exchange Commission (SEC) charged 27 firms and individuals with fraudulent promotion of stock on online platforms as reported by SEC (2017). Allegedly, authors were paid to promote penny stocks to pump the stocks' prices such that the originators can sell their stocks and profit, which is also known as a pump and dump scheme. Affected platforms include Seeking Alpha and Benzinga, which are also contained in the Nasdaq news data set. However, it is clear that this simple pump and dump strategy is harder to pull off for S&P 500 companies than for micro-cap stocks. As we consider only S&P 500 companies, this kind of fraudulent bullishness bias should not pose a problem. As for a possibly biased opinion of the authors, Zhang and Swanson (2010) found, that the opinions of day traders are overly optimistic. As numerous authors on platforms such as Seeking Alpha are indeed traders, this might also be a reason for the bullishness bias.

Figure 2 goes into more detail regarding the densities of the estimated sentiment. Here, we split the data by sector i and estimated sentiment L resulting in $n_{i,L}$ data points. Then we compute binned kernel density estimates (BKDE) while selecting the individual bandwidths $h_{i,L}$ with the oversmoothed bandwidth selector as described by Wand and Jones (1995). Furthermore, we estimate for each sector i and sentiment L its mean $\widehat{\mu}_{i,L}$ and standard deviation $\widehat{\sigma}_{i,L}$. Then we simulate $x_{i,L} = (x_{i,L,1}, x_{i,L,2}, \ldots, x_{i,L,n_{i,L}}) \sim \mathrm{N}(\widehat{\mu}_{i,L}, \widehat{\sigma}_{i,L})$ and estimate fit the BKDE again with bandwidth $h_{i,L}$.

As a result, the densities of BL and SM appear to be bimodal while this effect is not as strongly pronounced for the LM lexicon which might be because BL is more capable of capturing positive sentiment than LM as found by Zhang et al. (2016). Nonetheless, the estimated distributions for LM are also far from symmetrical around zero.

4 Panel Regression

Following Antweiler and Frank (2004), the effects of sentiment on the stock reactions (volatility and returns) are investigated by using contemporaneous regressions. Since Fama (1970), the efficient market hypothesis (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



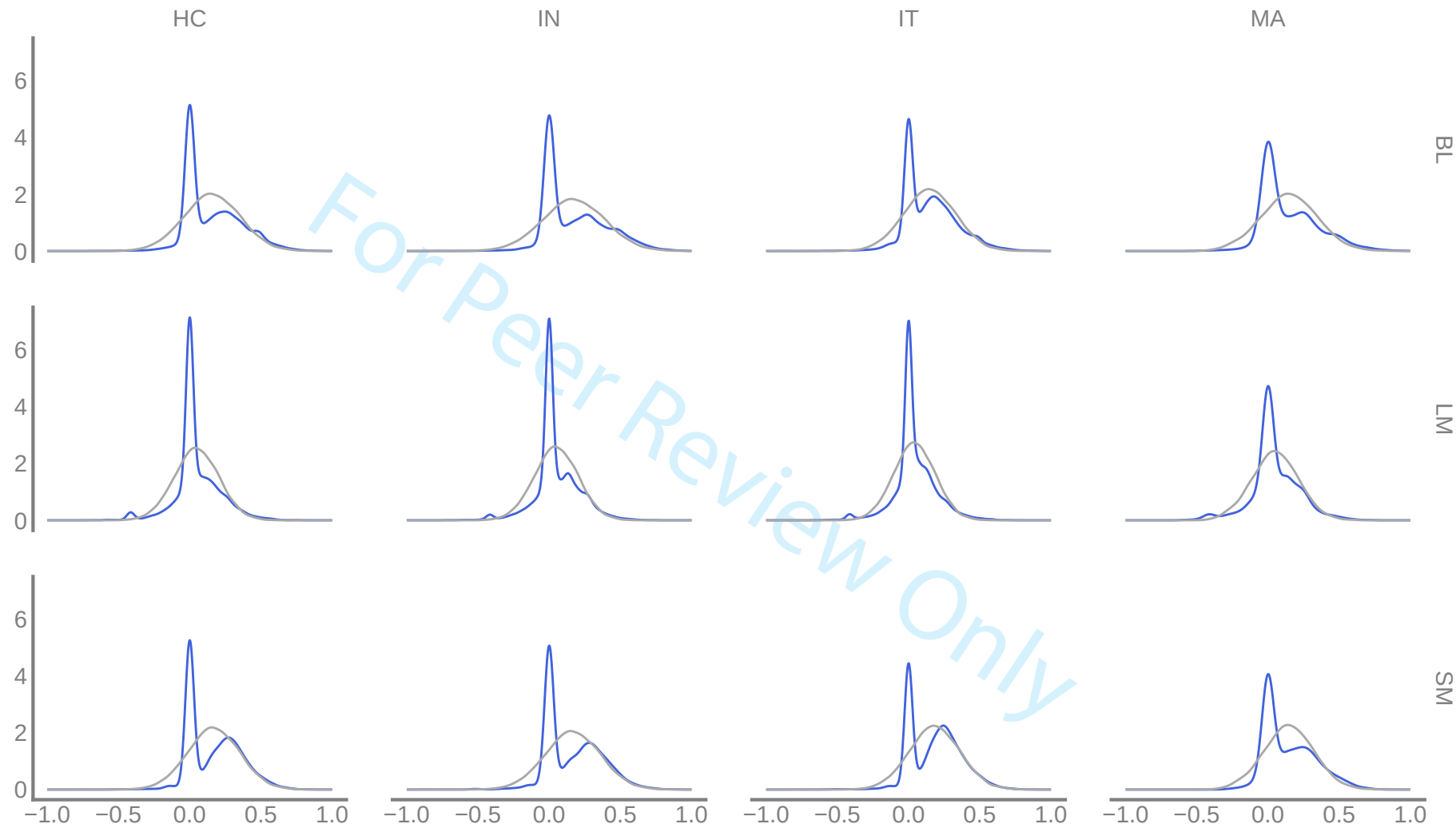


Figure 2: Estimated Densities of Sentiment by Sector and fitted Normal Distribution

information at each time point t. However, since the first mention of the EMH, it has been shown that markets are not necessarily 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

$$\sigma_{i,t} = \alpha_i + \beta_1^\top Sent_{i,t} + \beta_2^\top X_{i,t} + \gamma_i + \varepsilon_{i,t}, \quad (18)$$

$$r_{i,t} = \alpha_i + \beta_1^\top Sent_{i,t} + \beta_2^\top X_{i,t} + \gamma_i + \varepsilon_{i,t}. \quad (19)$$

As in Zhang et al. (2016), the models are estimated separately. γ_i corresponds to the fixed effect for firm i satisfying $\sum_i \gamma_i = 0$ and $\varepsilon_{i,t}$ is the error term of company i at day t. Recall that different 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.

$X_{j,t}$ 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 **??** and **??** we include only $R_{i,t}$ or $\sigma_{j,t}$, respectively. Both $\sigma_{j,t}$ and $R_{i,t}$ are included in equation **??**.

Sector	BL		LM		SM	
	$PF_{i,t}$	$PN_{i,t}$	$PF_{i,t}$	$PN_{i,t}$	$PF_{i,t}$	$PN_{i,t}$
	Panel A: Volatility $\log \sigma_{i,t}$					
Consumer Discretionary	-0.022 (0.022)	0.366^{***} (0.050)	0.034 (0.036)	0.360^{***} (0.058)	0.091^{***} (0.021)	0.695^{***} (0.073)
Consumer Staples	0.035 (0.029)	0.269^{***} (0.042)	0.125^{**} (0.053)	0.223^{***} (0.034)	0.120^{***} (0.034)	0.412^{***} (0.043)
Energy	0.057^{*} (0.030)	0.683^{***} (0.082)	0.367^{***} (0.049)	0.478^{***} (0.075)	0.201^{***} (0.035)	0.549^{***} (0.090)
Financials	0.037^{*} (0.021)	0.302^{***} (0.047)	0.127^{***} (0.037)	0.238^{***} (0.025)	0.176^{***} (0.037)	0.305^{***} (0.067)
Health Care	0.084^{***} (0.023)	0.384^{***} (0.051)	0.208^{***} (0.051)	0.359^{***} (0.053)	0.161^{***} (0.036)	0.827^{***} (0.097)
Industrials	-0.015 (0.024)	0.312^{***} (0.050)	0.039 (0.038)	0.242^{***} (0.037)	0.073^{**} (0.031)	0.483^{***} (0.084)
Information Technology	-0.037 (0.031)	0.304^{***} (0.044)	-0.062 (0.060)	0.384^{***} (0.044)	0.061^{**} (0.026)	0.609^{***} (0.073)
Materials	0.055^{**} (0.023)	0.450^{***} (0.087)	0.167^{***} (0.003)	0.346^{***} (0.026)	0.077^{*} (0.041)	0.647^{***} (0.171)
	Panel B: Returns $R_{i,t}$					
Consumer Discretionary	0.001^{***} (0.000)	-0.001 (0.001)	0.003^{***} (0.001)	-0.001 (0.001)	0.003^{***} (0.000)	-0.006^{**} (0.002)
Consumer Staples	0.001^{**} (0.000)	-0.000 (0.000)	0.002^{***} (0.001)	-0.001 (0.001)	0.001^{***} (0.000)	-0.004^{***} (0.001)
Energy	0.002^{**} (0.001)	-0.003^{**} (0.001)	0.002 (0.002)	-0.001 (0.002)	0.004^{***} (0.001)	-0.008^{**} (0.003)
Financials	0.001^{*} (0.000)	-0.003^{**} (0.001)	0.002^{**} (0.001)	-0.002^{*} (0.001)	0.003^{***} (0.001)	-0.008^{***} (0.002)
Health Care	0.001 (0.001)	-0.002 (0.001)	0.002^{*} (0.001)	-0.003^{***} (0.001)	0.001^{*} (0.000)	-0.003 (0.002)
Industrials	0.001^{***} (0.000)	-0.002^{**} (0.001)	0.003^{***} (0.000)	-0.002^{***} (0.001)	0.002^{***} (0.000)	-0.004^{***} (0.001)
Information Technology	0.002^{***} (0.001)	-0.002 (0.001)	0.003^{**} (0.001)	-0.003 (0.002)	0.003^{***} (0.001)	-0.004 (0.004)
Materials	0.002^{**} (0.001)	0.001^{***} (0.000)	0.003^{*} (0.002)	0.001 (0.001)	0.005^{***} (0.001)	-0.010^{***} (0.000)

Table 7: Contemporaneous Panel Regression (Fractions)

	BL				LM				SM			
Sector	$\lvert B_{i,t}\rvert$		$BN_{i,t}$		$\lvert B_{i,t}\rvert$		$BN_{i,t}$		$\lvert B_{i,t}\rvert$		$BN_{i,t}$	
	Panel C: Volatility $\log \sigma_{i,t}$											
Consumer Discretionary	−0.012	(0.022)	−0.321***	(0.087)	−0.045	(0.033)	−0.248***	(0.068)	0.082***	(0.019)	−0.239**	(0.103)
Consumer Staples	0.048	(0.036)	−0.145*	(0.077)	0.066	(0.051)	−0.056	(0.058)	0.125***	(0.037)	−0.164***	(0.056)
Energy	0.039	(0.046)	−0.524***	(0.123)	0.124**	(0.056)	−0.163***	(0.045)	0.146***	(0.054)	−0.146***	(0.054)
Financials	0.013	(0.032)	−0.139	(0.101)	−0.029	(0.056)	−0.141**	(0.056)	0.142***	(0.042)	0.043	(0.044)
Health Care	0.103***	(0.030)	−0.062	(0.090)	0.131***	(0.048)	−0.037	(0.048)	0.169***	(0.037)	−0.530***	(0.152)
Industrials	−0.006	(0.022)	−0.250***	(0.090)	−0.032	(0.027)	−0.134***	(0.047)	0.055*	(0.032)	−0.180**	(0.083)
Information Technology	−0.044	(0.030)	−0.131*	(0.079)	−0.138***	(0.048)	−0.330***	(0.080)	0.039	(0.029)	−0.284***	(0.082)
Materials	0.049***	(0.011)	−0.074	(0.147)	−0.007	(0.010)	−0.149**	(0.060)	0.098***	(0.027)	−0.675**	(0.314)
	Panel D: Returns $R_{i,t}$											
	$B_{i,t}$		$BN_{i,t}$		$B_{i,t}$		$BN_{i,t}$		$B_{i,t}$		$BN_{i,t}$	
Consumer Discretionary	0.001***	(0.000)	−0.001	(0.002)	0.002***	(0.001)	−0.001	(0.001)	0.003***	(0.000)	0.003	(0.003)
Consumer Staples	0.001*	(0.000)	−0.000	(0.001)	0.001	(0.001)	0.001	(0.001)	0.001***	(0.000)	−0.000	(0.003)
Energy	0.001	(0.001)	0.003	(0.004)	0.002*	(0.001)	−0.002	(0.002)	0.003***	(0.001)	0.006	(0.006)
Financials	0.001**	(0.000)	0.001	(0.002)	0.001	(0.001)	0.001	(0.001)	0.002***	(0.001)	0.001	(0.003)
Health Care	0.001	(0.001)	−0.003	(0.003)	0.001	(0.001)	0.002**	(0.001)	0.001**	(0.000)	−0.002	(0.004)
Industrials	0.001***	(0.000)	0.003	(0.002)	0.002***	(0.000)	0.001	(0.001)	0.002***	(0.000)	0.000	(0.001)
Information Technology	0.002**	(0.001)	0.001	(0.003)	0.001	(0.001)	0.004**	(0.002)	0.002***	(0.001)	−0.001	(0.006)
Materials	0.002**	(0.001)	0.001***	(0.000)	0.003*	(0.001)	−0.005***	(0.002)	0.004***	(0.000)	0.008	(0.008)

Table 8: Contemporaneous Panel Regression (Bullishness)

Sector	BL		LM		SM	
	$PF_{i,t}$	$PN_{i,t}$	$PF_{i,t}$	$PN_{i,t}$	$PF_{i,t}$	$PN_{i,t}$
	Panel E: Volatility $\log \sigma_{i,t+1}$					
Consumer Discretionary	-0.050^{***} (0.015)	0.148^{***} (0.033)	-0.045^{*} (0.024)	0.114^{***} (0.029)	0.052^{***} (0.016)	0.233^{***} (0.042)
Consumer Staples	-0.047^{**} (0.023)	0.173^{***} (0.032)	0.021 (0.020)	0.079^{**} (0.032)	0.086^{***} (0.020)	0.143^{**} (0.058)
Energy	-0.014 (0.015)	0.326^{***} (0.026)	0.109^{***} (0.022)	0.235^{***} (0.039)	0.110^{***} (0.023)	0.211^{***} (0.056)
Financials	-0.016 (0.025)	0.071^{**} (0.030)	-0.048 (0.043)	0.102^{***} (0.019)	0.123^{***} (0.034)	0.007 (0.035)
Health Care	0.021 (0.018)	0.155^{***} (0.052)	0.082^{**} (0.033)	0.107^{**} (0.050)	0.134^{***} (0.025)	0.202^{***} (0.046)
Industrials	-0.008 (0.019)	0.093^{***} (0.023)	0.030 (0.034)	0.034^{*} (0.020)	0.124^{***} (0.023)	0.068 (0.065)
Information Technology	-0.077^{***} (0.021)	0.105^{**} (0.045)	-0.085^{*} (0.044)	0.095^{***} (0.035)	0.055^{***} (0.021)	0.124^{***} (0.048)
Materials	-0.084^{***} (0.009)	0.371^{***} (0.009)	-0.150^{***} (0.004)	0.295^{***} (0.039)	0.119^{***} (0.044)	0.149^{**} (0.073)
	Panel F: Returns $R_{i,t+1}$					
Consumer Discretionary	-0.000 (0.001)	-0.000 (0.001)	0.001 (0.001)	-0.001 (0.002)	0.001^{*} (0.001)	-0.003 (0.002)
Consumer Staples	0.000 (0.001)	0.001 (0.001)	0.002^{**} (0.001)	-0.000 (0.001)	0.001^{***} (0.000)	-0.002^{*} (0.001)
Energy	0.003^{***} (0.001)	-0.000 (0.002)	0.004^{***} (0.002)	0.001 (0.002)	0.002^{**} (0.001)	0.003 (0.003)
Financials	-0.000 (0.001)	0.001 (0.001)	0.003^{*} (0.001)	-0.000 (0.001)	0.002^{*} (0.001)	-0.001 (0.001)
Health Care	-0.001 (0.001)	-0.000 (0.001)	-0.003^{***} (0.001)	0.001 (0.001)	-0.002^{**} (0.001)	0.000 (0.001)
Industrials	-0.001^{**} (0.000)	0.000 (0.001)	-0.001 (0.001)	-0.000 (0.001)	0.001 (0.001)	-0.000 (0.001)
Information Technology	0.000 (0.001)	-0.003^{*} (0.002)	0.000 (0.001)	-0.003^{**} (0.001)	-0.001 (0.001)	-0.003^{*} (0.002)
Materials	-0.003 (0.002)	-0.000^{*} (0.000)	-0.006^{***} (0.002)	-0.000 (0.001)	-0.005^{***} (0.001)	0.010^{***} (0.002)

Table 9: Lagged Panel Regression (Fractions)

Sector	$\lvert B_{i,t-1}\rvert$		$BN_{i,t-1}$		$\lvert B_{i,t-1}\rvert$		$BN_{i,t-1}$		$\lvert B_{i,t-1}\rvert$		$BN_{i,t-1}$	
	Panel G: Volatility $\log \sigma_{i,t+1}$											
Consumer Discretionary	-0.077^{***}	(0.020)	-0.407^{***}	(0.061)	-0.116^{***}	(0.029)	-0.261^{***}	(0.064)	0.044^{*}	(0.022)	-0.358^{***}	(0.106)
Consumer Staples	-0.042	(0.029)	-0.424^{***}	(0.112)	-0.038	(0.035)	-0.105^{**}	(0.044)	0.096^{***}	(0.026)	-0.150^{*}	(0.081)
Energy	-0.079	(0.053)	-0.865^{***}	(0.147)	-0.057	(0.058)	-0.536^{***}	(0.103)	0.073	(0.061)	-0.464^{***}	(0.102)
Financials	-0.075^{***}	(0.029)	-0.264^{**}	(0.116)	-0.190^{***}	(0.057)	-0.388^{***}	(0.061)	0.060	(0.045)	0.146^{**}	(0.057)
Health Care	0.041	(0.025)	-0.115	(0.113)	0.044	(0.044)	-0.039	(0.051)	0.159^{***}	(0.033)	-0.424^{***}	(0.145)
Industrials	-0.015	(0.022)	-0.570^{***}	(0.116)	-0.048	(0.036)	-0.220^{***}	(0.041)	0.086^{***}	(0.023)	-0.118	(0.075)
Information Technology	-0.127^{***}	(0.025)	-0.337^{***}	(0.095)	-0.190^{***}	(0.046)	-0.330^{***}	(0.085)	-0.014	(0.030)	-0.139^{**}	(0.066)
Materials	-0.030^{***}	(0.005)	-0.858^{***}	(0.079)	-0.131^{***}	(0.038)	-0.637^{***}	(0.000)	0.123^{**}	(0.048)	-0.451^{***}	(0.014)
	Panel H: Returns $R_{i,t}$											
	$B_{i,t}$		$BN_{i,t}$		$B_{i,t}$		$BN_{i,t}$		$B_{i,t}$		$BN_{i,t}$	
Consumer Discretionary	-0.000	(0.001)	0.004	(0.002)	0.000	(0.001)	0.001	(0.001)	0.001	(0.001)	0.011	(0.007)
Consumer Staples	0.000	(0.000)	0.004	(0.003)	0.000	(0.001)	0.002	(0.001)	0.001^{***}	(0.000)	0.004^{*}	(0.002)
Energy	0.002^{***}	(0.001)	0.000	(0.006)	0.002^{*}	(0.001)	-0.002	(0.003)	0.003^{*}	(0.001)	-0.009	(0.006)
Financials	0.000	(0.001)	-0.004	(0.003)	0.002^{*}	(0.001)	-0.001	(0.002)	0.002^{**}	(0.001)	-0.005^{***}	(0.002)
Health Care	-0.001	(0.001)	0.003	(0.004)	-0.002^{*}	(0.001)	0.001	(0.002)	-0.002^{**}	(0.001)	0.005	(0.003)
Industrials	-0.001^{**}	(0.000)	-0.001	(0.003)	-0.001	(0.001)	0.000	(0.001)	0.001	(0.001)	0.001	(0.002)
Information Technology	-0.000	(0.001)	0.007^{*}	(0.004)	-0.000	(0.001)	0.003	(0.002)	-0.001	(0.001)	0.010^{**}	(0.004)
Materials	-0.003	(0.002)	-0.001	(0.002)	-0.004^{***}	(0.001)	0.003^{***}	(0.000)	-0.004^{***}	(0.001)	-0.011	(0.010)

Table 10: Lagged Panel Regression (Bullishness)

Sector	BL		LM		SM	
	Panel G: Volatility $\log \sigma_{i,t+1}$					
Consumer Discretionary	-0.023^{***}	(0.008)	-0.025^{***}	(0.010)	-0.029^{***}	(0.008)
Consumer Staples	-0.024^{***}	(0.009)	-0.029^{**}	(0.011)	-0.037^{***}	(0.010)
Energy	-0.068^{***}	(0.009)	-0.087^{***}	(0.010)	-0.072^{***}	(0.009)
Financials	-0.021	(0.014)	-0.023	(0.015)	-0.032^{**}	(0.014)
Health Care	-0.051^{***}	(0.010)	-0.054^{***}	(0.010)	-0.060^{***}	(0.011)
Industrials	-0.028^{***}	(0.010)	-0.031^{***}	(0.010)	-0.031^{***}	(0.010)
Information Technology	-0.008	(0.011)	-0.008	(0.015)	-0.019^{*}	(0.011)
Materials	-0.027^{***}	(0.004)	-0.032^{***}	(0.008)	-0.035^{***}	(0.003)
	Panel H: Returns $R_{i,t+1}$					
Consumer Discretionary	0.000	(0.000)	0.000	(0.000)	0.000	(0.000)
Consumer Staples	-0.000^{**}	(0.000)	-0.000	(0.000)	-0.000^{***}	(0.000)
Energy	-0.001	(0.000)	-0.000	(0.000)	-0.001	(0.001)
Financials	-0.000	(0.000)	-0.000	(0.000)	-0.000	(0.000)
Health Care	0.001^{**}	(0.000)	0.001^{**}	(0.000)	0.001^{***}	(0.000)
Industrials	0.000	(0.000)	-0.000	(0.000)	-0.000	(0.000)
Information Technology	0.000	(0.000)	0.000	(0.000)	0.001	(0.001)
Materials	0.002^{*}	(0.001)	0.001^{***}	(0.000)	0.002^{***}	(0.001)

Table 11: Lagged Panel Regression (Bullishness)

5 Appendix

References

Al-Rjoub, S. A. M. (2016). Media Coverage of Central Bank Communications and Stock Market Reactions. *SSRN Electronic Journal*.

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

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python*. O'Reilly, Beijing ; Cambridge [Mass.], 1st ed edition. OCLC: ocn301885973.

Cao, H. H., Coval, J. D., and Hirshleifer, D. A. (2001). Sidelined investors, trading-generated news, and security returns. *Dice Working Paper No. 2000-2*.

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

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. and Chen, M. (2007). Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science*, pages 1375–1388.

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.

Guo, L., Huang, D., Tu, J., and Wang, R. (2017). Is news informative or sentimental to analysts?

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. *10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004)*, pages 168–177.

Härdle, W., Lee, Y.-J., Schäfer, D., and Yeh, Y.-R. (2009). Variable selection and oversampling in the use of smooth support vector machines for predicting the default risk of companies. *Journal of Forecasting*, 28(6):512–534.

Jurafsky, D. and Martin, J. H. (2009). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education Internat, Upper Saddle River, NJ, 2. ed., pearson internat. ed edition. OCLC: 263455133.

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

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65.

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, 1(4):309–317.

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.

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., and Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts: Good Debt or Bad Debt. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Polanyi, L. and Zaenen, A. (2006). *Contextual Valence Shifters*, pages 1–10. Springer Netherlands, Dordrecht.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Rifkin, R. and Klautau, A. (2004). In defense of one-vs-all classification. *J. Mach. Learn. Res.*, 5:101–141.

Rönnqvist, S. and Sarlin, P. (2017). Bank distress in the news: Describing events through deep learning. *Neurocomputing*.

Rosa, C. and Verga, G. (2007). On the consistency and effectiveness of central bank communication: Evidence from the ecb. *European Journal of Political Economy*, 23(1):146–175.

Schumaker, R. P., Zhang, Y., Huang, C.-N., and Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems*, 53(3):458–464.

SEC (2017). Payments for Bullish Articles on Stocks Must Be Disclosed to Investors. https://www.sec.gov/news/press-release/2017-79.

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

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Wand, M. P. and Jones, M. C. (1995). *Kernel smoothing*. Number 60 in Monographs on statistics and applied probability. Chapman & Hall, London ; New York, 1st ed edition.

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

Wiebe, J. and Riloff, E. (2005). *Creating Subjective and Objective Sentence Classifiers from Unannotated Texts*, pages 486–497. Springer Berlin Heidelberg, Berlin, Heidelberg.

Wilson, T. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *In Proceedings of HLT-EMNLP*, pages 347–354.

Zhang, J. L., Härdle, W. K., Chen, C. Y., and Bommes, E. (2016). Distillation of news flow into analysis of stock reactions. *Journal of Business & Economic Statistics*, 34(4):547–563.

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

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Sector	Short	Code	Industry Groups	Subcode
Energy	EN	10	Energy	1010
Materials	MAT	15	Materials	1510
Industrials	IND	20	Capital Goods	2010
			Commercial & Professional Services	2020
			Transportation	2030
Consumer	CD	25	Automobiles & Components	2510
Discretionary			Consumer Durables & Apparel	2520
			Consumer Services	2530
			Media	2540
			Retailing	2550
Consumer Staples	CS	30	Food & Staples Retailing	3010
			Food, Beverage & Tobacco	3020
			Household & Personal Products	3030
Health Care	HC	35	Health Care Equipment & Services	3510
			Pharmaceuticals, Biotechnology & Life Sciences	3520
Financials	FIN	40	Banks	4010
			Diversified Financials	4020
			Insurance	4030
			Real Estate	4040
Information	IT	45	Software & Services	4510
Technology			Technology Hardware & Equipment	4520
			Semiconductors & Semiconductor Equipment	4530
Telecommunication	TS	50	Telecommunication Services	5010
Services				
Utilities	UT	55	Utilities	5510

Table 12: GICS Sector Specification

IRTG 1792 Discussion Paper Series 2018

For a complete list of Discussion Papers published, please visit irtg1792.hu-berlin.de.

001 "Data Driven Value-at-Risk Forecasting using a SVR-GARCH-KDE Hybrid" by Marius Lux, Wolfgang Karl Härdle and Stefan Lessmann, January 2018.

002 "Nonparametric Variable Selection and Its Application to Additive Models" by Zheng-Hui Feng, Lu Lin, Ruo-Qing Zhu asnd Li-Xing Zhu, January 2018.

003 "Systemic Risk in Global Volatility Spillover Networks: Evidence from Option-implied Volatility Indices " by Zihui Yang and Yinggang Zhou, January 2018.

004 "Pricing Cryptocurrency options: the case of CRIX and Bitcoin" by Cathy YH Chen, Wolfgang Karl Härdle, Ai Jun Hou and Weining Wang, January 2018.

005 "Testing for bubbles in cryptocurrencies with time-varying volatility" by Christian M. Hafner, January 2018.

006 "A Note on Cryptocurrencies and Currency Competition" by Anna Almosova, January 2018.

007 "Knowing me, knowing you: inventor mobility and the formation of technology-oriented alliances" by Stefan Wagner and Martin C. Goossen, February 2018.

008 "A Monetary Model of Blockchain" by Anna Almosova, February 2018.

009 "Deregulated day-ahead electricity markets in Southeast Europe: Price forecasting and comparative structural analysis" by Antanina Hryshchuk, Stefan Lessmann, February 2018.

010 "How Sensitive are Tail-related Risk Measures in a Contamination Neighbourhood?" by Wolfgang Karl Härdle, Chengxiu Ling, February 2018.

011 "How to Measure a Performance of a Collaborative Research Centre" by Alona Zharova, Janine Tellinger-Rice, Wolfgang Karl Härdle, February 2018.

012 "Targeting customers for profit: An ensemble learning framework to support marketing decision making" by Stefan Lessmann, Kristof Coussement, Koen W. De Bock, Johannes Haupt, February 2018.

013 "Improving Crime Count Forecasts Using Twitter and Taxi Data" by Lara Vomfell, Wolfgang Karl Härdle, Stefan Lessmann, February 2018.

014 "Price Discovery on Bitcoin Markets" by Paolo Pagnottoni, Dirk G. Baur, Thomas Dimpfl, March 2018.

015 "Bitcoin is not the New Gold - A Comparison of Volatility, Correlation, and Portfolio Performance" by Tony Klein, Hien Pham Thu, Thomas Walther, March 2018.

016 "Time-varying Limit Order Book Networks" by Wolfgang Karl Härdle, Shi Chen, Chong Liang, Melanie Schienle, April 2018.

017 "Regularization Approach for NetworkModeling of German EnergyMarket" by Shi Chen, Wolfgang Karl Härdle, Brenda López Cabrera, May 2018.

018 "Adaptive Nonparametric Clustering" by Kirill Efimov, Larisa Adamyan, Vladimir Spokoiny, May 2018.

019 "Lasso, knockoff and Gaussian covariates: a comparison" by Laurie Davies, May 2018.

IRTG 1792, Spandauer Straße 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.

IRTG 1792 Discussion Paper Series 2018

For a complete list of Discussion Papers published, please visit irtg1792.hu-berlin.de.

020 "A Regime Shift Model with Nonparametric Switching Mechanism" by Haiqiang Chen, Yingxing Li, Ming Lin and Yanli Zhu, May 2018.

021 "LASSO-Driven Inference in Time and Space" by Victor Chernozhukov, Wolfgang K. Härdle, Chen Huang, Weining Wang, June 2018.

022 " Learning from Errors: The case of monetary and fiscal policy regimes" by Andreas Tryphonides, June 2018.

023 "Textual Sentiment, Option Characteristics, and Stock Return Predictability" by Cathy Yi-Hsuan Chen, Matthias R. Fengler, Wolfgang Karl Härdle, Yanchu Liu, June 2018.

024 "Bootstrap Confidence Sets For Spectral Projectors Of Sample Covariance" by A. Naumov, V. Spokoiny, V. Ulyanov, June 2018.

025 "Construction of Non-asymptotic Confidence Sets in 2 -Wasserstein Space" by Johannes Ebert, Vladimir Spokoiny, Alexandra Suvorikova, June 2018.

026 "Large ball probabilities, Gaussian comparison and anti-concentration" by Friedrich Götze, Alexey Naumov, Vladimir Spokoiny, Vladimir Ulyanov, June 2018.

027 "Bayesian inference for spectral projectors of covariance matrix" by Igor Silin, Vladimir Spokoiny, June 2018.

028 "Toolbox: Gaussian comparison on Eucledian balls" by Andzhey Koziuk, Vladimir Spokoiny, June 2018.

029 "Pointwise adaptation via stagewise aggregation of local estimates for multiclass classification" by Nikita Puchkin, Vladimir Spokoiny, June 2018.

030 "Gaussian Process Forecast with multidimensional distributional entries" by Francois Bachoc, Alexandra Suvorikova, Jean-Michel Loubes, Vladimir Spokoiny, June 2018.

031 "Instrumental variables regression" by Andzhey Koziuk, Vladimir Spokoiny, June 2018.

032 "Understanding Latent Group Structure of Cryptocurrencies Market: A Dynamic Network Perspective" by Li Guo, Yubo Tao and Wolfgang Karl Härdle, July 2018.

033 "Optimal contracts under competition when uncertainty from adverse selection and moral hazard are present" by Natalie Packham, August 2018.

034 "A factor-model approach for correlation scenarios and correlation stress-testing" by Natalie Packham and Fabian Woebbeking, August 2018.

035 "Correlation Under Stress In Normal Variance Mixture Models" by Michael Kalkbrener and Natalie Packham, August 2018.

036 "Model risk of contingent claims" by Nils Detering and Natalie Packham, August 2018.

037 "Default probabilities and default correlations under stress" by Natalie Packham, Michael Kalkbrener and Ludger Overbeck, August 2018.

038 "Tail-Risk Protection Trading Strategies" by Natalie Packham, Jochen Papenbrock, Peter Schwendner and Fabian Woebbeking, August 2018.

IRTG 1792, Spandauer Straße 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.

IRTG 1792 Discussion Paper Series 2018

For a complete list of Discussion Papers published, please visit irtg1792.hu-berlin.de.

039 "Penalized Adaptive Forecasting with Large Information Sets and Structural Changes" by Lenka Zbonakova, Xinjue Li and Wolfgang Karl Härdle, August 2018.

040 "Complete Convergence and Complete Moment Convergence for Maximal Weighted Sums of Extended Negatively Dependent Random Variables" by Ji Gao YAN, August 2018.

041 "On complete convergence in Marcinkiewicz-Zygmund type SLLN for random variables" by Anna Kuczmaszewska and Ji Gao YAN, August 2018.

042 "On Complete Convergence in Marcinkiewicz-Zygmund Type SLLN for END Random Variables and its Applications" by Ji Gao YAN, August 2018.

043 "Textual Sentiment and Sector specific reaction" by Elisabeth Bommes, Cathy Yi-Hsuan Chen and Wolfgang Karl Härdle, September 2018.

IRTG 1792, Spandauer Straße 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.