Chen, Cathy Yi-Hsuan; Fengler, Matthias R.; Härdle, Wolfgang Karl; Liu, Yanchu

Working Paper

Textual Sentiment, Option Characteristics, and Stock Return Predictability

IRTG 1792 Discussion Paper, No. 2018-023

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

Suggested Citation: Chen, Cathy Yi-Hsuan; Fengler, Matthias R.; Härdle, Wolfgang Karl; Liu, Yanchu (2018) : Textual Sentiment, Option Characteristics, and Stock Return Predictability, IRTG 1792 Discussion Paper, No. 2018-023, Humboldt-Universität zu Berlin, International Research Training Group 1792 "High Dimensional Nonstationary Time Series", Berlin

This Version is available at:
https://hdl.handle.net/10419/230734

IRTG 1792 Discussion Paper 2018-023

Textual Sentiment, Option Characteristics, and Stock Return Predictability

Cathy Yi-Hsuan Chen *
Matthias R. Fengler *[2]
Wolfgang Karl Härdle *
Yanchu Liu *[3]

* Humboldt-Universität zu Berlin, Germany
*[2] University of St. Gallen, Switzerland
*[3] Sun Yat-sen University, Guangzhou,

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

Textual Sentiment, Option Characteristics, and Stock Return Predictability

Cathy Yi-Hsuan Chen, [*] Matthias R. Fengler, [†] Wolfgang Karl Härdle, [‡]
Yanchu Liu [§]

June 18, 2018

Abstract

We distill sentiment from a huge assortment of NASDAQ news articles by means of machine learning methods and examine its predictive power in single-stock option markets and equity markets. We provide evidence that single-stock options react to contemporaneous sentiment. Next, examining return predictability, we discover that while option variables indeed predict stock returns, sentiment variables add further informational content. In fact, both in a regression and a trading context, option variables orthogonalized to public and sentimental news are even more informative predictors of stock returns. Distinguishing further between overnight and trading-time news, we find the first to be more informative. From a statistical topic model, we uncover that this is attributable to the differing thematic coverage of the alternate archives. Finally, we show that sentiment disagreement commands a strong positive risk premium above and beyond market volatility and that lagged returns predict future returns in concentrated sentiment environments.

Key words: investor disagreement; option markets; overnight information; stock return predictability; textual sentiment; topic model; trading-time information;

JEL Classification: C58, G12, G14, G41

[*]C.A.S.E.- Center for Applied Statistics & Economics, Humboldt-Universität zu Berlin, Spandauer Str. 1, 10178 Berlin, Germany (*e-mail: cathy.chen@hu-berlin.de*)

[†]University of St. Gallen, School of Economics and Political Science, Bodanstrasse 6, CH-9000 St. Gallen, Switzerland (*e-mail: matthias.fengler@unisg.ch*)

[‡]C.A.S.E.- Center for Applied Statistics & Economics, Humboldt-Universität zu Berlin and Sim Kee Boon Institute, School of Business, Singapore Management University, 50 Stamford Road, Singapore 178899 and Xiamen University, Xiamen, China (*e-mail: haerdle@hu-berlin.de*)

[§]Department of Finance, Lingnan (University) College, Sun Yat-sen University, Haizhu District, Guangzhou, China, 510275 (*e-mail: liuych26@mail.sysu.edu.cn*)

1 Introduction

News moves the stock market. While this hypothesis underlies most models of asset pricing, only in recent years has research gotten to the actual heart of this statement. Based on large bodies of text, there is now growing evidence that news, or more precisely, the sentimental tone expressed via the news items, carries informational content for price discovery in equity markets that extends beyond the information sets created from past observations and other traditional market variables, such as the Fama French factors (Antweiler and Frank, 2004; Tetlock, 2007, 2010; Cujean and Hasler, 2016; Bommes et al., 2018). Separate from this strand, several studies address the predictive role of option price data for stock markets (Dennis and Mayhew, 2002; Pan and Poteshman, 2006; Xing et al., 2010; Stilger et al., 2016). Here the predictive power is attributed to the notion that informed traders maximize the value of their private information about stocks by trading in the option market. Leverage and fewer market frictions, as imposed, e.g., by short-sell constraints, create attractive trading incentives and therefore induce demand for particular option contracts, which in turn leads to their predictive content about future asset prices.

How do we accommodate these different narratives of asset pricing? Clearly, apart from private information, investors also derive their outlook for a particular stock partly from public information, such as news or analysts' reports, and could, as they increase their familiarity with it, choose the option market as their preferred marketplace. Indeed, Han (2008) reports that sentiment is a driver of variation in index option prices. One may therefore conjecture that sentiment influences the equity market and the option market alike and hence the decision to trade in the option market relies upon a mixture of both private and public information. Consequently, it is desirable to separate the sentimental component from the private information embedded in option price data.

In this work, we study the entire nexus of textual sentiment, option data characteristics, and stock return predictability. We employ advanced text analytic tools based on supervised learning methods to distill firm-level sentiment from a large text corpus scraped from NASDAQ news

feed channels pertaining to 97 companies of the S&P100 index. In a first step, we analyze how trading-hour sentiment impacts three key single-stock option data characteristics, namely implied volatility, out-of-the-money put prices, and the implied volatility skew. We establish that both firm-level sentiment as well as the cross-sectional aggregates of firm-level sentiments, i.e., sentiment indices, have a measurable impact on these option data characteristics. This augments the observations of Han (2008) made for S&P500 index options.

With this empirical evidence at hand, we examine the predictive power of single-stock option characteristics (OCs) for S&P100 equity returns. In line with previous research, we find that OCs predict stock returns. Remarkably, they continue to do so in the presence of sentiment variables, whereby the negative sentiment index emerges as a particularly powerful predictor variable. To study this predictive power more closely, we use the sentiment data along with supplementary traditional predictor information to extract the purported private content of option data. Using these orthogonalized components of OCs, which we obtain by regressing OCs on a set of sentiment variables and control variables, we find that they still predict stock return data and do so more precisely. In order to check the economic significance of the statistical results, we compare the profits of two trading strategies, where the first is based on OCs only, while the second one builds on the orthogonalized OCs. We find that the latter strategy dominates the former in terms of Sharpe ratio, no matter which OC it is based on. We conclude that (1) both private and public information is absorbed in option data; (2) the amount of private information about stocks intrinsic to option data is substantial; (3) a trading strategy based on approximative private information after filtering out the public fraction of sentiment achieves a higher profitability than one that does not partial out public sentiment.

We then study the role of sentiment dispersion for stock return predictability. In doing so, we exploit the fact that the cross-sectional distribution of firm-level sentiment yields a natural measure of sentimental agreement over the firms included in the panel. From a theoretical perspective, it has been debated as early as Miller (1977) whether investor disagreement triggers lower stock prices, the rationale being that if pessimists stay out of the market because of short-sale constraints, asset prices reflect only the optimists' price appraisals and hence are overvalued. Al-

ternatively, it has been suggested by Varian (1985), David (2008), Cujean and Hasler (2016) and others that disagreement should be related to higher future stock prices because disagreement gives rise to a risk factor which investors ask to be compensated for. In our empirical assessment, we find that investors' sentimental disagreement gives rise to a risk premium above and beyond standard market volatility risk. Because sentimental disagreement is only slightly if at all correlated with market return volatility, we take this as strong support for Varian's risk premium hypothesis.

In a final step, we explore price reversals and momentum patterns conditional on states of sentimental consensus. To this end, we consider the intersection between sentimental disagreement and the lagged returns. In concentrated states of sentiment, we find that both low and high returns tend to be followed by low returns, implicating price reversal on positive returns and momentum on negative returns. This gives a new sentimental twist to the role of disagreement for return prediction, because the disagreement measures we use are not constructed from analysts' reports as is common practice (Banerjee, 2011; Yu, 2011; Kim et al., 2014).

In this work, we also discover new results about the dissimilar informational content of trading-hour versus overnight information. In fact, all our predictive stock return regressions point toward overnight information, i.e., information collected from articles in the night preceding (not overlapping) the return measurement, which is more informative than the "younger" trading-time sentiment, i.e., information collected during the last trading time. This is an unanticipated finding, as one may expect the overnight sentiment to be fully absorbed in prices during the following trading session. In order to obtain a better understanding of this phenomenon, we apply a statistical topic model to the two alternate archives of news. We find that while trading-time and overnight articles share similar topics related to dividends and earnings, they vary in terms of emphasis as regards the remaining topics. Overnight articles of our text corpus tend to focus on fundamental aspects of the investment strategy, for instance, by featuring topics like economic outlook and general investment strategies, whereas trading-time articles lean toward tactical topics such as trading signals obtained from capital movements of funds and, most interestingly, trading opportunities via the option market. These differing emphases, in connection with less

complex topics being dealt with during trading-time, may contribute to the distinct predictive power of the different news archives. We thus corroborate observations about the relevance of overnight information in other fields such as accounting (Berkman and Truong, 2009; Doyle and Magilke, 2009), market micro structure (Barclay and Hendershott, 2003; Moshirian et al., 2012), and realized variance prediction (Wang et al., 2015; Buncic and Gisler, 2016), albeit from a very different angle.

As regards our techniques of sentiment extraction, we build on a more refined tool kit than traditionally used in the extant literature. Usually, based on a "bag-of-words" document model, one employs a dictionary-based counting process after natural language pre-processing, which involves stemming and lemmatization. To create text-based sentiments, these unsupervised learning methods are used, for instance, in Cao et al. (2002), Das and Chen (2007), Schumaker et al. (2012), Chen et al. (2014), and Zhang et al. (2016), building on dictionaries, such as that of Loughran and McDonald (2011), among others. Challenging the popularity of lexicon projection, Bommes et al. (2018) observe that supervised learning algorithms trained on the financial phrase bank of Malo et al. (2014) for sentence-based sentiment extraction realize far superior classification results because they accomplish a surpassing comprehension of the linguistic sentence structure. Following these insights, we therefore use a supervised learning algorithm trained on this particular phrase bank as our foremost tool to predict sentence-level sentiment, but keep all computations for sentiment variables which are derived from a traditional lexicon projection based on the Loughran-McDonald lexicon for robustness purposes.

The outline of this work is as follows: In Section 2, we present the techniques used to quantify sentiments, deferring discussion of details to the appendix. Section 3 describes the text corpus and option data, and how we define firm-level and market-level sentiment measures. We study sentiment and option data in Section 4. Section 5 researches stock return predictability and Section 6 studies sentimental disagreement. Section 7 concludes.

2 Sentiment quantification

This section describes our methods to quantify sentiment qualitatively; more details are given in the appendix. We pursue two strategies: a classical lexicon or "bag-of-words" approach and a refined supervised learning method based on a linear scoring function. Both methods allows us to construct a firm-level sentiment quantification, which we call "bullishness." The algorithms were programmed in Python and R and the natural language processing was carried out with the Python module "Natural Language Processing Toolkit" of Bird et al. (2009). The algorithms are available as quantlets on quantlet.de

2.1 Lexicon method (LM)

Lexicon-based sentiment extraction is a widely applied technique in text analytics. It is based on a "bag-of-words" model for a document and works by projecting into a predefined dictionary, i.e., by counting positive, negative, or neutral words. Weighting and averaging yields a fraction of positive (negative) words per day per document, where the term "document" can refer to a whole article or any substructure, such as a sentence. Our dictionary of choice is the Loughran and McDonald (2011) lexicon as it has been developed on purpose to parse financial news and is also a fundamental tool in, e.g., Thompson Reuters financial services.

While this word-based approach is widely used, it has been argued that sentiment measured on the sentence level describes the investors' mood more precisely, because it is expected to have a better semantic orientation than the pure "bag-of-words" approach (Wiebe and Riloff, 2005; Wilson et al., 2005). We therefore aggregate the sentence-based polarity over all sentences of an article to a fraction of total negative and positive polarity of each company and day; see Eqs. (9) and (10) in the appendix.

The fraction of polarity words is used, e.g., by Chen et al. (2014) and Zhang et al. (2016) as a measure of sentiment, whereas Antweiler and Frank (2004) go one step further to combine both negative and positive sentiment into a single measure of bullishness. Following these ideas, we

specify

$$B_{i,t} = \frac{\log(1+FP_{i,t}) - \log(1+FN_{i,t})}{\log(2)} \tag{1}$$

as our measure of bullishness for company i on day t. One can easily observe that $B_{i,t} < 0$ holds if the polarity of the text is relatively negative, while $B_{i,t} = 0$ indicates neutrality and $B_{i,t} > 0$ suggests a positive polarity. Eq. (1) defines the bullishness for a given document. If a firm i is referred to in more than one document on date t we compute multiple $B_{i,t}$ and average them.

2.2 Supervised method (SM)

As an alternative to the simple lexical projections of dictionary elements and their refinements based on contextual polarity, we looked into a supervised learning approach; see Malo et al. (2014). They investigate how semantic orientations can be detected in financial and economic news by looking at the overall sentence structure. To this end, they established a human-annotated finance phrase-bank, which enhances a basic financial lexicon by incorporating contextual semantic orientations in financial and economic news texts. On this training data set we train a score-based linear discrete response model of the form $s(X) = \beta^\top X$, where $\beta \in \mathbb{R}^p$ is a parameter vector and has possibly a large dimension p. After comparing various classification loss functions and penalties, we estimate the prediction model based on the hinge loss and the L_1 penalty.

The mean accuracy of the SM sentence-level method (with oversampling) is 80%, whereas the one based on the LM lexical projection achieves only an accuracy of 64%. A deeper analysis through the confusion matrix, which we report in Table 1, reveals that LM more often produces false negatives (type 2 error) and false positives (type 1 error) than the SM method does. For the case of True $= -1$, we calculate the false positive rates of SM and LM as 0.21 (the ratio of 289+254 to 2535) and 0.58 (the ratio of 289+12 to 514), respectively. The false negative rate of SM and LM are, respectively, 0.09 (the ratio of 96+105 to 2184) and 0.59 (the ratio of 200+111 to 524). Obviously, the SM with oversampling achieves higher precision (equivalent to 1−type 2 error) and higher recall (equivalent to 1−type 1 error). In sum, SM is better at returning more relevant results (recall), and more relevant results than the irrelevant ones (precision).

From training, we obtain a huge vector $\widehat{\beta}$ with dimension $p \approx 43500$ which enters the score $s(X) = \widehat{\beta}^\top X$. To predict sentiment, $\widehat{\beta}$ is applied to the NASDAQ article database. Each document is split up into its sentences and the corresponding score is calculated, yielding a predictor for the polarity, which then leads to analogues of (9) and (10), and finally (1); see the appendix for more details. As a result we obtain the bullishness $B_{i,t}$ for each document, company and day of our sample period.

3 Data

3.1 Text corpus

We consider news articles that are available through the NASDAQ news platform, which were written between Jan. 1 2012 to Apr. 30 2016 by professional reporters and analysts. NASDAQ offers a platform for news and financial articles from selected contributors including leading media such as Reuters, MT Newswires, RTT news, or investment research firms such as Motley Fool, Zacks or GuraFocus. The news contents is classified into a number of categories, e.g., stocks, economy, world news, politics, commodities, technology, and fundamental analysis. News in the stocks category accounts for a big proportion with the symbols assigned by NYSE, NASDAQ, or other exchanges. The time stamp, the date, the contributor, the symbols, the title, and the complete text are all extracted via an automatic web scraper written by Zhang et al. (2016) and extended to the more recent period in this research. It is available for academic purposes at the Research Data Center (RDC) at the Humboldt-Universität zu Berlin. It should be noted that while the data origin suggests that only companies traded on the NASDAQ are covered, articles about companies listed at other exchanges are available too.

In total, we find 344 631 articles over this period. Reducing the data set to articles about at least one company listed in a pool of 97 firms across 9 industry sectors, all of which are constituents of S&P 100, leaves us with 119 680 articles; see the appendix for the complete list.[1] The number

[1] AbbVie Inc. (ABBV) is the only firm that is covered as of Jan. 2013.

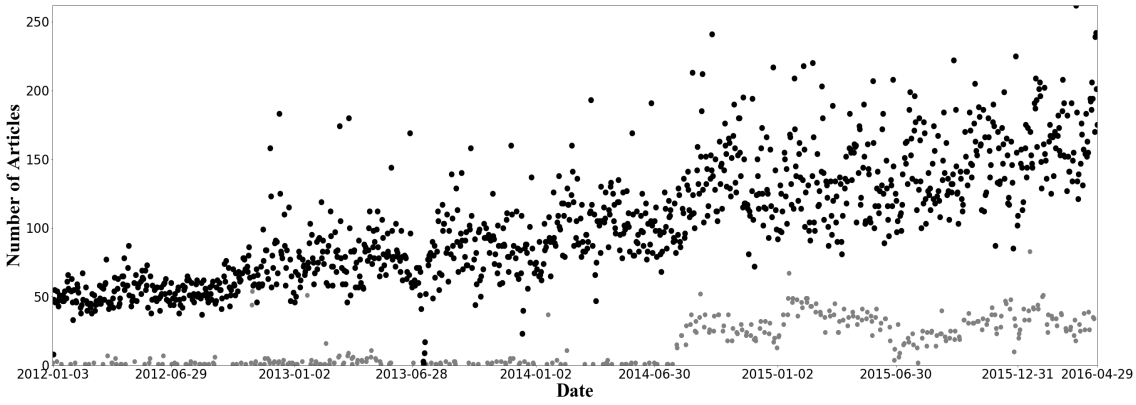

Figure 1: Number of article postings per day referring to the 97 companies listed in the S&P 100 index. A black point indicates the number of articles posted on a trading day, a gray point the number of articles posted on a non-trading day (weekend, holiday).

of firms we can make use of for this study is limited by the availability of single-stock option data on the firms covered by NASDAQ articles (see also Section 3.3). In total, the sample period contains 1 581 calendar days, out of which 1 088 are trading days. Thus, the 97 firms are receivers of approximately one piece of news per day. Figure 1 illustrates the number of published articles per day over the sample period. Articles posted on trading days are more numerous than those released on non-trading days (weekends, holidays). One can also observe a positive linear trend in the number of articles posted on trading days and a jump in the number of postings on non-trading days after Jun. 30 2014, possibly due to an increasing popularity of the NASDAQ news platform over time.

113 080 (94.49%) out of the 119 680 articles are posted on trading days. To further exhibit the intraday news posting activity during trading days, we display in Figure 2 a histogram on an hourly scale, based on the time stamps of all trading-day articles (black dots in Figure 1). The trading hours on NYSE and NASDAQ are from 09:30:00 a.m. to 03:59:59 p.m. Eastern time. The period from 00:00:00 a.m. to 09:29:59 a.m. and that from 04:00:00 p.m. to 11:59:59 p.m. on each trading day are called non-trading hours. Figure 2 reveals a number of noteworthy patterns about the posting behavior. There are 33 160 articles (29.32%) posted before market opening at 09:30:00 a.m., most of which (20 821 articles or 18.4%) appear during the half hour before market opening (i.e., between 09:00:00 a.m. and 9:29:59 a.m.). This observation coincides with

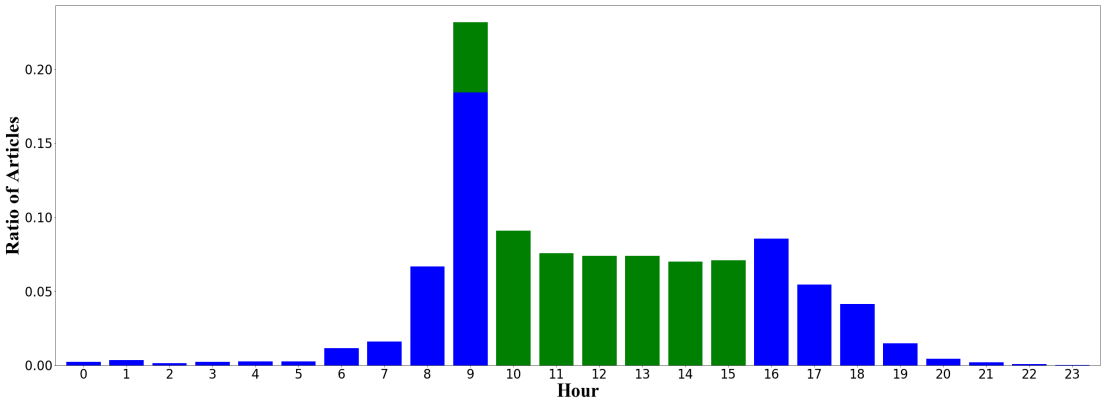

Figure 2: Hourly distribution (ET) of NASDAQ article postings. Hourly labels indicate the full hour, say, from 08:00:00 a.m. to 08:59:59 a.m., etc. Blue indicates non-trading hours, green trading hours. Height of bar denotes the frequency of articles posted during that hour. The hour from 9:00:00 a.m. to 9:59:59 a.m. is split into two parts due to market opening at 09:30:00 a.m. The histogram is computed only from postings on trading days (black dots in Figure 1).

the tradition of morning conferences within the finance industry. Financial news reporters and analysts usually send out a large number of reports and prospectuses for the market and equities to their customers immediately after the morning conferences. Moreover, there are 56 833 articles (50.26%) posted in an almost even fashion during the trading hours. The sample documents 23 087 articles (20.42%) after 04:00:00 p.m., most of which are posted before 07:00:00 p.m. After 07:00:00 p.m., the number of article postings subsides and remains low till about 06:00:00 a.m. Thus, most article posting is concentrated during typical working hours.

The fact that about half of the trading day articles are posted when markets are closed (and more than one half, when adding on top the articles posted on weekends and holidays) motivates us to investigate the relationship between the news items' topics and their posting times. For this purpose, we employ a topic model on each set of articles (trading-time versus overnight articles, including weekends and holidays). This statistical topic model allows us to discover the hidden thematic structures in the two news archives. The specific model we use is a Latent Dirichlet Allocation (LDA), which builds on a "bag-of-words" approach to text data and allows each article to have multiple topics, while the overall number of topics over the entire archive is constant and fixed by the researcher. The LDA uses the joint distribution over the observed (the words in the articles) and the hidden random variables (the latent topics defined as a distribution over sets

of words) to compute the conditional distribution of the hidden topic structure conditional on observed words. From the collection of the most frequent words for each topic, one can infer its thematic content; for more details, we refer to Blei (2012) and Linton et al. (2017).

We display the results of the LDA fits in Tables 2 and 3, which report the top 15 most frequent words over the selected 10 topics. Conspicuously, the topic structures vary between trading time and overnight postings, both in terms of their content and their order of occurrence. As regards the overnight topics in Table 2, we find *Dividends, Investment Strategies, Earnings, Equities, Asset Management, Global Outlook, Charts Analysis, Analyst Roundups, Sector Analysis* and *Market*. Among the trading-time articles we uncover *Press releases, Earnings 1, Funds, Option trades, Charts, Sectors, Dividends, Equities, Earnings 2*, and *Share types*.[2]

Comparing both topic collections, we observe that while some topics of general significance to investors (*Dividends, Chart Analysis, Earnings, Sectors, Press releases/Analyst roundups*) are common across the alternate news archives, although at different orders of importance, we can identify topics which are markedly distinct between them. Besides this, the overnight archive tends to cover basic principles of strategic asset allocation, such as investing in growth, momentum, value stocks (topic 2), general stock coverage (topic 4), and the global economic outlook (topic 6). In contrast, the trading-time articles appear to feature tactical aspects like trading signals or trading opportunities. More specifically, we find *Funds*, which discusses capital inflows and outflows into and from ETFs, possibly as relevant trading indicators of the state of the market; and *option trades* (topic 4), a topic of obvious interest for this research, which exhibits words like `options, trading, using` and expiry dates as `october, january, november`. Moreover, the 19th top word, not shown in Table 3, is `yieldboost`, which further underpins our interpretation of this topic theme. These observations are insightful for interpreting our later results. In anticipation of these, we find that news covered in articles posted during trading time impacts the contemporaneous option variables; however, in the predictive stock return regressions, we observe that the content of articles posted during market close is more informative than that of articles posted during

[2] The two earnings topics differ in terms of emphasis: *Earnings 1* discusses surprise elements of earnings statements featuring words like `miss, beat, surprise, estimate`, whereas *Earnings 2*, with words like `indicator, history, reaction, sensitive` appears to offer a broader discussion of the theme.

trading hours.

3.2 Sentiment measures

After applying the sentiment quantification methods as described in Sections 2.1 and 2.2, we obtain two firm-specific bullishness measures for each trading day: a trading-hour measure $B_{i,t}$ and an overnight measure $B^{on}_{i,t}$. The time index t is defined as follows: For a trading day t at NYSE, the trading hour period is from 09:30:00 a.m. to 03:59:59 p.m. in New York time (GMT−5); the overnight period indexed with t is from 04:00:00 p.m. at $t-1$ and 09:29:59 a.m. on date t. For this reason, trading sentiment on t is more recent than overnight sentiment on t. Moreover, for a trading day on a Friday, the overnight sentiment will also cover the entire weekend till the morning of the next trading date. This design helps align the date structure between the textual news channel and the option trading data. Note that this definition of non-trading time differs from the one applied to compute the histogram in Figure 2.

Time aggregation to trading days and matching with the option data yields a final sample size of 105 283 daily firm-specific sentiment values. In summary, we study the following sentiment variables:

(1) firm-specific bullishness $B_{i,t}$ ($B^{on}_{i,t}$) for the trading hour period (the overnight period): positive value of $B_{i,t}$ or $B^{on}_{i,t}$ implies positive sentiment and vice versa;

(2) firm-specific negative bullishness defined as $BN_{i,t} = -B_{i,t}\,\mathbf{I}(B_{i,t} < 0)$ for the trading hour period (accordingly $BN^{on}_{i,t}$ for the overnight period);

(3) an aggregate sentiment index $B_{idx,t}$ ($B^{on}_{idx,t}$) for the trading hour period (the overnight period) as an equally weighted cross-sectional average of $B_{i,t}$ ($B^{on}_{i,t}$);

(4) an aggregate negative sentiment index $BN_{idx,t}$ ($BN^{on}_{idx,t}$) for the trading hour period (the overnight period), as an equally weighted cross-sectional average of the $BN_{i,t}$ ($BN^{on}_{i,t}$).

We compute the aggregate sentiment indices because firm-specific bullishness may carry senti-

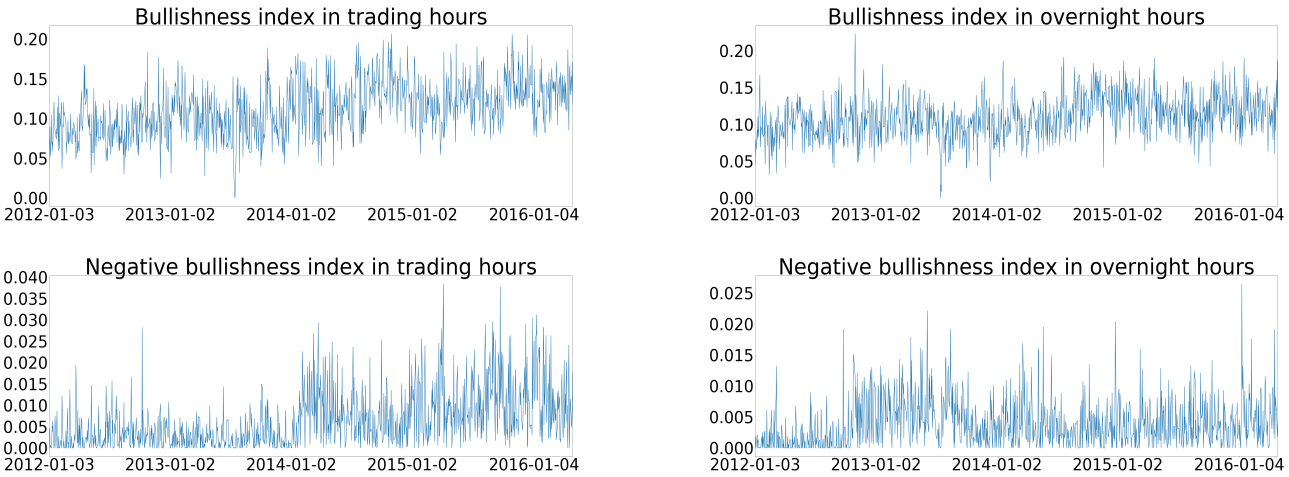

Figure 3: Daily bullishness index B_{idx} and B^{on}_{idx}, and the negative bullishness index BN_{idx} and BN^{on}_{idx}, constructed during the trading hours (left-hand panel) and the overnight (right-hand panel), are displayed. Underlying sentiment is derived from the SM method.

mental content that is informative for other firms. For illustration, Figure 3 exhibits the time series evolution of the (SM-based) daily bullishness index B_{idx} and the negative bullishness index BN_{idx} that we obtain from the NASDAQ article text corpus.

It should be noted that the market-wide indices we construct from firm-level sentiment differ in nature from concurrent sentiment indices, such as the market-based sentiment index of Baker and Wurgler (2006), the survey-based University of Michigan Consumer Sentiment Index, and a search-based index as in Da et al. (2014). In constructing the indices, we exclusively rely on the text-based information of the NASDAQ articles. In contrast, as criticized by Sibley et al. (2016), the widely used Baker and Wurgler (2006) sentiment index is mostly made up of other risk factors, such as stock market conditions and the business cycle in general. On the other hand, we are not compelled to identify the relevant sentiment-revealing search terms, such as `recession`, `bankruptcy`, or `unemployment`, as in Da et al. (2014), which could bias actual market sentiment.

Aside from the cross-sectional average, we also study cross-sectional dispersion and its impact on asset returns. In Figure 4, we display a Gaussian kernel density fit of $B_{i,t}$ on a selected number of dates. The days are chosen between Jan. 2012 and April 2016 for each half year to exhibit the evolution of the cross-sectional sentiment over time. We observe that sentiment clusters around

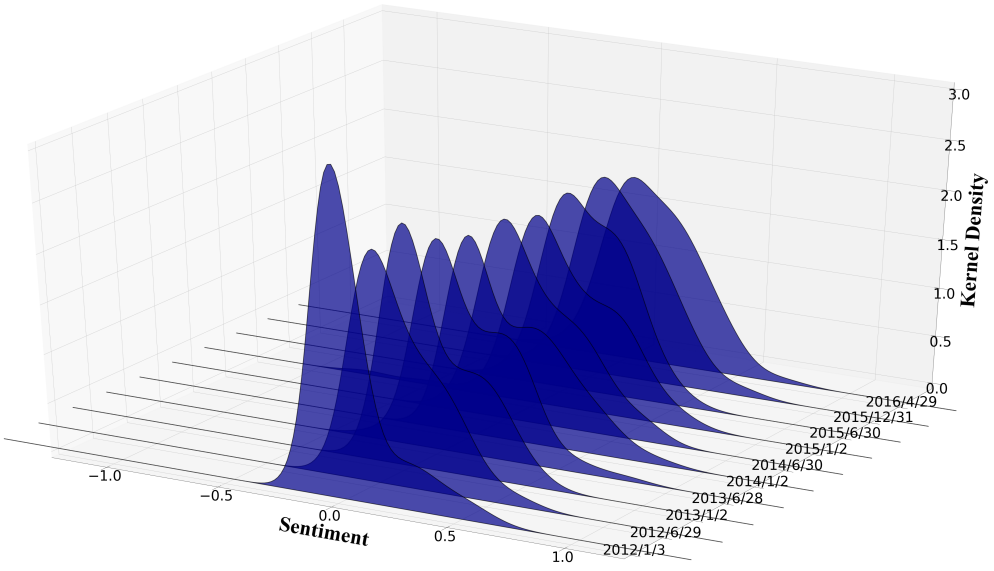

Figure 4: Cross-sectional density evolution of bullishness B_i over time, based on the SM method. Gaussian kernel density estimates for each half year of the sample period.

zero, implying that many firms get neutral coverage or no articles; second, we diagnose variation with times of lean and peaked, or dispersed and skewed densities.

Summary statistics of the sentiment data over all 97 firms are displayed in the upper panel of Table 4. Three important observations can be made. First, from the 25% quantiles of BN_i and BN_i^{on}, it can be inferred that negative news is much more rare than positive news in our sample. In part, this may be related to our sample ranging from Jan. 2012 to Apr. 2016; however, it is also known that negative views are generally less likely to be expressed than positive ones. In the sentiment construction, we account for this fact by oversampling; see Bommes et al. (2018). Second, the statistical properties of sentiment gathered from the articles either during a trading day or overnight are qualitatively similar. Our empirical analysis will investigate whether the two data sources are also similar in terms of economic content. Third, comparing LM-based sentiment projections with those obtained from SM, we find a larger mean for SM compared to LM, whereas standard deviations are of similar size. Thus, LM sentiments exhibit a much larger variation

relative to their mean than do SM sentiments. A higher variation of LM-based sentiment might be attributed to its "bag-of-words" nature. The "bag-of-words" model is insensitive to word order and grammar and therefore features no understanding of language structure. As a consequence, a sentiment tone produced with alternative words, but having the same sentimental intention, can produce very different sentiment scores. This effect results in a larger variation of the scores.

In Table 5, we study the correlation structure between the sentiment variables and the option characteristics (OC) data; see Section 3.3 for details on OC data. For the 97 sample firms, a cross-sectional correlation between any pair of OC of firm i and the sentiment variables at the firm-level or market-level is documented for the 25%, 50% and 75% quantile values. One can observe that the $B_{i,t}$ distilled from SM display a higher negative correlation with $Skew_{i,t}$ than the one extracted from the LM. The same observation is evident not only for the firm-level sentiment but also for the market-level sentiment $B_{idx,t}$. For the other two OCs, $IV_{i,t}$ and $Put_{i,t}$, the discrepancies in correlations are even more manifest. Looking at the 25% quantile of correlations, we observe that $B_{idx,t}$ obtained from SM is more negatively associated with $IV_{i,t}$ and $Put_{i,t}$ than the corresponding $B_{idx,t}$ measure obtained for LM; on the other hand, $BN_{idx,t}$ computed by SM is more positively associated with $IV_{i,t}$ and $Put_{i,t}$ than the measure obtained by means of LM. In summary, the informational content of the sentiment quantified by means of SM, in comparison to that of LM, is more accordant with that of OCs.

We also check the correlation of the bullishness index and the negative bullishness index with other major market factors such as the market, market volatility and Fama-French factors. Both measures hardly correlate with them. This suggests that these indices capture factors largely orthogonal to commonly accepted market factors.

3.3 Option and stock market data

We match daily stock and option data to the text corpus. More specifically, we collect end-of-day total return data, bid and ask option price quotes, and implied volatility (IV) data from the IvyDB US database offered by OptionMetrics. As additional controls, we merge daily Fama-French 5-

factor data collected from Kenneth R. French's website[3] to the data set.

The option characteristics (OC) used are defined as follows:

- $Skew_{i,t}$: volume-weighted average IV of out-the-money (OTM) put options minus volume-weighted average IV of at-the-money (ATM) call options at time t of firm i;
- $Put_{i,t} = \log(1 + p_{i,t})$: where $p_{i,t}$ is the mid price (average price of best bid and best offer) of the available OTM put prices for each trading day t, weighted by trading volumes and divided by spot price;
- $IV_{i,t}$: volume-weighted average of IV of the available ATM options on each trading day.

Moneyness, throughout this paper, is defined as the ratio of the strike price to the stock price. OTM is defined as moneyness between 0.80 and 0.95; ATM is moneyness between 0.95 and 1.05. To ensure sufficient liquidity, the options with time-to-maturities between 10 and 60 days are included. Summary statistics of the OC data over all 97 firms are displayed in the lower part of Table 4.

4 The predictive content of textual sentiment

4.1 Equity option data and investor sentiment

How do option markets react to sentiment? The first to address this question on a market-wide level was Han (2008). This author finds evidence suggesting that the IV smile of S&P 500 options is steeper (flatter) when market sentiment is more bearish (bullish). In that study, market-wide sentiment of institutional investors is measured by the proportion of bullish investors minus the proportion of bearish investors based on newsletter surveys done by *Investors Intelligence*. Our study differs on two accounts. First, instead of using newsletter survey data to measure sentiment,

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

we quantify news-based sentiment through proper textual analysis. Second, we work with a firm-specific sentiment quantification in a panel data framework to study the common effect in the cross-sectional data set.

The underlying rationale of our approach is that the market participants reading the NASDAQ articles can choose a marketplace to implement a trading idea inspired by the perceived article's sentiment. The marketplace can be either the stock market or the option market or both. Accordingly, the news quantified by sentiment impacts stock and option markets alike, but possibly with different speeds of dissemination. Dennis and Mayhew (2002), Chakravarty et al. (2004) and, more recently, Xing et al. (2010) claim that trading can be accomplished in an easier and more cost-efficient fashion if trades are executed via the option markets, e.g., by selling calls or buying puts, rather than on the stock market. With this rationale in mind, we formulate

Hypothesis 1 (H1): Firm-level option characteristics reflect firm-specific sentiment.

As set out in Section 3.3, we employ the option characteristics (OC) $Skew_{i,t}$, $Put_{i,t}$ and $IV_{i,t}$ as sensors of option market reactions. We check these three OCs as dependent variables in the fixed-effects regressions:

$$OC_{i,t} = \alpha + \gamma_i + \beta_1 B_{i,t} + \beta_2^\top X_t + \varepsilon_{i,t}\,, \tag{2}$$

where $\{Skew_{i,t}, Put_{i,t}, IV_{i,t}\} \in OC_{i,t}$, $B_{i,t}$ is the quantified trading-time bullishness of firm i at time t, see (1). Finally, X_t is the vector of control variables including the Fama-French five factors and the stock return, its volatility and market volatility.

In (2) a potential endogeneity problem may exist. This is because the NASDAQ article might not be the original source of a specific piece of news. Although the majority of articles are released before the closing time of option markets (4 p.m. ET); see Figure 2, orthogonality of $\varepsilon_{i,t}$ and $B_{i,t}$ requires that the article in the NASDAQ platform be the exclusive source of a particular piece of news. This could be challenged, because rather than representing original news, an article could have been written in response to a press release of a referenced company earlier in the day. Indeed, we find exogeneity formally rejected using standard endogeneity tests of the Hausman-Wu type. Therefore, we treat $B_{i,t}$ as an endogenous regressor in (2) and use the lagged sentiment $B_{i,t-1}$ as

a natural instrument for $B_{i,t}$. Appropriately, the regression results of Table 6 are obtained from a two-stage instrumental variable regression.

As can be inferred from Table 6, H1 is strongly supported in the presence of all controls. We find that $B_{i,t}$ is significantly related to $Skew_{i,t}$, $Put_{i,t}$ and $IV_{i,t}$. As negative news is released and bearish sentiment is formed subsequently, investors may want to engage in long positions in put options, resulting in a rising price of OTM put options. As a consequence, the IV of OTM puts over the IV of ATM calls, namely the volatility skew, is expected to rise. In addition to the risk on the downside, i.e., $Skew_{i,t}$ and $Put_{i,t}$, the benchmark variance risk proxied by IV of ATM options shows an opposite response: lower sentiment means higher IV, i.e., ATM IV declines on positive news.

The results support H1 and further corroborate the findings of Han (2008) in that firm-level sentiment impacts single-stock option prices. In addition, our evidence emphasizes the price discovery role of option markets. The ability of price discovery is subject to the market design, which comprises an array of market microstructure features. Chakravarty et al. (2004) ascribe the price discovery role of option markets to leverage and built-in downside risk. Due to these features, both informed and uninformed traders have incentives to trade in this marketplace. This research documents this fact by quantifying the impact of news on option prices. Moreover, in Section 4.2, we distinguish between the informational content of OCs as reflected by sentiment, i.e., a public part, and a residual component, which captures private information.

Given the empirically established relation between firm-level OCs and firm-level sentiment, one may ask whether individual OCs react to the content of aggregate news. In addition to the firm-level sentiment, we conjecture that the OCs react to aggregate sentiment, which represents the common or systematic sentiment component in the text corpus:

Hypothesis 2 (H2): Firm-level option characteristics reflect aggregate sentiment.

H2 can be cast into the regression

$$OC_{i,t} = \alpha + \gamma_i + \beta_1 B_{i,t} + \beta_2 B_{idx,t} + \beta_3 BN_{idx,t} + \beta_4^\top X_t + \varepsilon_{i,t} \tag{3}$$

where $\{Skew_{i,t}, Put_{i,t}, IV_{i,t}\} \in OC_{i,t}$, and $B_{idx,t}$ is the trading-time sentiment index and $BN_{idx,t}$ is the trading-time negative sentiment index as introduced in Section 3.2.

As shown in Table 6, the aggregate sentiment index provides incremental information on option markets of S&P100 companies. In the presence of higher negative market sentiment $BN_{idx,t}$, we see a higher volatility skew, higher OTM prices and higher ATM implied volatility; by contrast, we observe the reverse response with rising market bullishness $B_{idx,t}$. Remarkably, firm-level sentiment remains significant despite the presence of market-wide sentiment. Looking at Table 7, where sentiment is discovered by the LM method, we find additional support for these results as far as OTM prices and IV is concerned. For the skew, results are only weakly supported, or as in the case of BN_{idx}, defy expectations. Recalling that type 2 error rates of lexicon projection are high for negative statements (see Section 2.2), we do not overinterpret this counterintuitive result.

4.2 Equity return predictability of option characteristics

A growing body of literature attributes a prominent role for the derivatives market to price discovery in spot markets; see, e.g., Chakravarty et al. (2004), Pan and Poteshman (2006), Chang et al. (2013), and Conrad et al. (2013). In particular, Xing et al. (2010) show that option characteristics, such as *Skew*, predict the cross-sectional distribution of stock returns. The authors hypothesize that this is so because traders possessing a private information advantage over the public execute their trading ideas in the option market and subsequently profit from it as their private information diffuses in the market. In their study, the private information is related to future firm fundamentals.

Given the evidence provided in Table 6, however, a natural question is to what extent, if any, traders actually act on private information. It could well be that trading ideas, which are inspired by the sentiment articulated in the NASDAQ articles, are executed via the option market. For this reason, we include both option characteristics and sentiment variables together in our predictive regressions of stock returns. If option characteristics are no longer significant with public sentiment

being controlled for, we may discount the importance of inside information implied in option characteristics. We therefore build the following hypothesis:

H3: Besides private information, sentiment contributes to stock return predictability.

We explore this question by means of the regression equation

$$R_{i,t+1} = \alpha + \theta^\top \mathbf{B}_t + \gamma OC_{i,t} + \beta^\top X_{i,t} + \varepsilon_{i,t} \tag{4}$$

where $R_{i,t+1}$ denotes the return of firm i at time $t+1$ and $\{Skew_{i,t}, Put_{i,t}, IV_{i,t}\} \in OC_{i,t}$. $\mathbf{B}_t$ is a vector of sentiment-related variables including $B_{i,t}$, $B_{idx,t}$,$BN_{idx,t}$, $B^{on}_{i,t}$, $B^{on}_{idx,t}$ and $BN^{on}_{idx,t}$.

In Table 8, we first report in scenarios (1) to (3) the results without sentiment. They all confirm the evidence of Xing et al. (2010): the volatility skew marginally predicts future returns, while the negative sign shows that the volatility skew is a signal of future stock underperformance (Stilger et al., 2016). Scenarios (2) and (3) show that OTM put and IV are both significantly positive. Thus, both OCs carry the undertone of a risk premium in the sense of the risk-return trade-off relation. In order to induce investors to hold assets when either volatility risk (IV) or downside risk (OTM put) is high, assets must offer a risk premium as compensation. These findings are widely confirmed in the literature (Bollerslev et al., 2013; Chen et al., 2018)

In scenarios (4) to (6), we include the sentiment information obtained from the NASDAQ articles as distilled by the SM method. As is apparent, firm-level sentiment $B_{i,t}$ is insignificant, which is consistent with Tetlock (2007), Stambaugh et al. (2012), and Zhang et al. (2016). In contrast, the negative trading-hour bullishness index has a clear directional impact on next day's returns: the higher is $BN_{idx,t}$, the lower the future return. For the bullishness index $B_{idx,t}$, which includes both positive and negative sentiment, no prediction power is found. Thus, the prediction power between average market-wide and negative market-wide sentiment is asymmetric and return prediction is only achievable in the presence of negative market sentiment. Theoretically, predictability in states of low market sentiment can stem from short-sale constraints, which defer trading (Diamond and Verrecchia, 1987; Engelberg et al., 2012). Expensive or prohibited short-selling of stocks reduces the speed of adjustment of security prices to private information, and thus leads to return

predictability.

We additionally investigate the predictive role of overnight sentiment. We find – as with trading-hour firm-level sentiment – no predictive power in firm-level overnight sentiment; the market-wide variables $B^{on}_{idx,t}$ and $BN^{on}_{idx,t}$, however, do carry significant predictive power. Thus, in comparison to trading-time information $B_{idx,t}$, there emerges an informational wedge between the sentiment indices of the alternate news archives. Whereas both negative indices and $B^{on}_{idx,t}$ provide predictive content, $B_{idx,t}$ does not. It is challenging, however, to ascertain where this informational wedge ensues from. As discussed in Section 3.3, the archives have a differing emphasis in terms of topics. The overnight archive offers more fundamental and strategic discussions, while the trading-time archive tends to feature tactical aspects of trading; such a discrepancy could contribute to the informational wedge. On the other hand, it could be that overnight information is generally more fundamental and hence more relevant or simply deals with more complex issues. Indicative of this presumption is the order of the respective topics within the archives; see again Tables 2 and 3. The first four topics in the overnight archive are *Dividends, Investment strategies, Earnings* and *Equities*, which are fundamentally important topics; in contrast, the first four topics in the trading-time archive are press releases/analyst blogs, (surprise elements of) earning reports, capital movements within and out of funds and option trading, all of which appear to be of more short-term interest. Indeed, the notion that more complex information requires time to be absorbed by the market and therefore is strategically placed during market close is a common thread in the accounting literature; see, e.g., Berkman and Truong (2009) and Doyle and Magilke (2009). Likewise, it has been observed that information that is impounded during non-trading hours and then reflected in the overnight return is crucial for accurately predicting future realized variance (Wang et al., 2015; Buncic and Gisler, 2016).

In scenarios (7) to (9), we report the results for sentiment variables based on LM. Overall, they support the previously discussed findings, with two key differences. First, it appears that firm-level $B_{i,t}$ negatively (and marginally) predicts the next day's return, which could be interpreted as an overreaction of stock returns to firm-level sentiment. While one could rationalize such overreactions in behavioral models of trading (Antweiler and Frank, 2004), we are cautious about such

an interpretation. Indeed, comparing the classification results of the SM method with those of the LM method points in a much different direction. As discussed in Section 2.2, in interpreting the confusion matrix in Table 1, the LM method is prone to producing many false negative classifications. In particular for negative sentiment ($True = -1$), we find about 60% false negatives, which is the largest type 2 error overall. Hence, the sentiment extracted from the LM method tends to be biased towards an overly pessimistic scale, which can explain the seeming overreaction reaction patterns documented in Table 8. As a second difference, the market-wide negative overnight sentiment $BN^{on}_{idx,t}$ has no predictive power. Because the negative sentiment index accumulates the aforementioned false negatives, it appears tempting to attribute the inferior informativeness to the very same cause. The remaining results are fully supported.

Across all scenarios, the conclusions as regards OTM put and IV as regressors remain the same when sentiment-related variables are included. Summing up, we find strong support of H3.

5 Sources of predictability: information advantage or sentiment?

In view of our findings in Section 4, we now isolate the purported private information component in OCs and provide statistical and economic evidence of its existence.

5.1 Regression results

The existing literature supports price discovery in option markets because private information about stock fundamentals is exploited via the option market. However, one could question whether the predictability stemming from trading on private information can be attributed entirely to private information. It is possible that the option market serves as a vehicle to quickly trade on public information. The results in Table 6 are supportive of this conjecture. Here, we carry out an anatomy of the "information content of option characteristics" and study to what extent the

predictability stems from an information advantage or needs to be ascribed to a certain preference of a marketplace. In short, we have

Hypothesis 4 (H4): OCs orthogonal to sentiment are informative about future stock returns.

H4 is concerned with the question of whether the public information as distilled in the sentiment score $B_{i,t}$ absorbs the predictive power of OCs for future returns. This is checked by the panel regression (5) that incorporates the residuals of the $OC_{i,t}$ regressed on the sentiment variables. By partialling out the public information and therefore operating on information orthogonal to sentiment-related information, we touch upon the fraction of unobserved information driving future returns. More precisely, we run the regressions

$$R_{i,t+1} = \alpha + \theta^\top \mathbf{B}_t + \gamma OC_{i,t}^{\perp} + \beta^\top X_{i,t} + \varepsilon_{i,t} \tag{5}$$

where $\mathbf{B}_t$ is a vector of sentiment-related variables including $B_{i,t}$, $B_{idx,t}$,$BN_{idx,t}$, $B_{i,t}^{on}$, $B_{idx,t}^{on}$ and $BN_{idx,t}^{on}$. $\{Skew_{i,t}^{\perp}, Put_{i,t}^{\perp}, IV_{i,t}^{\perp}\} \in OC_{i,t}^{\perp}$. $Skew_{i,t}^{\perp}$ is estimated as the residuals by regressing $Skew_{i,t}$ on $\mathbf{B}_t$ and control variables $X_{i,t}$. Likewise, $Put_{i,t}^{\perp}$ and $IV_{i,t}^{\perp}$ are estimated in the same way. $Skew_{i,t}^{\perp}$, $Put_{i,t}^{\perp}$ and $IV_{i,t}^{\perp}$ are orthogonal to public information and adjusted for a market-wide risk premium.

Table 9 shows the evidence for all scenarios discussed in Table 8. Picking scenarios (1), (4) and (7) as examples, $Skew_{i,t}^{\perp}$ corrected for public information enters into the equations with a negative coefficient. In fact, the OCs orthogonalized to sentiment appear to be more precise measures of information: p-values drop to about 5% as opposed to 10% as before in Table 8. In all other dimensions, the results are almost identical to those reported previously.

We summarize a number of implications. First, public information-adjusted OCs predict future returns and tend to do so more precisely; second, the market-wide sentiment is informative, but the firm-level sentiment is not. We may therefore conclude that the return predictability of OCs can be attributed to these two sources: (i) the market-relevant sentiment; and (ii) private information.

5.2 Private information long-short trading strategy

To further investigate the economic significance of private information reflected in the OC-residuals $OC_{i,t}^{\perp}$, we design a long-short trading strategy. Indeed, if the $OC_{i,t}^{\perp}$ is an isolated component of private information, it seems reasonable to expect a trading strategy based on $OC_{i,t}^{\perp}$ alone to be superior than to based directly on $OC_{i,t}$.

We execute the trading strategies on daily data. For any trading day t in the period from January 02, 2015 to April 29, 2016, the portfolio is constructed by the following steps:

Step 1: Compute the OC-residuals for each firm on day t, from the regression of the OC on the sentiment variables and the control variables as outlined in the previous section **(e.g. in (3))**. We use an in-sample period with three years before day t to calibrate the coefficients of the regression equations.

Step 2: Sort the 97 firms on day t in descending order of the residuals and separate them into deciles. If OC is $Skew$ (IV or Put), we sell (buy) the group with the highest residuals and sell (buy) the group with the lowest residuals, with equal weights.[4]

Step 3: Proceed to day $t+1$, calculate the return of the long-short portfolio, and rebalance. The three-year in-sample training period to determine regression coefficients is rolled forward.

We compare our strategy with the purely OC-based strategy of Xing et al. (2010). The latter is constructed in that one uses the day t's OCs to sort the 97 firms and builds up a long-short portfolio for the day $t+1$ similar to the one in Step 2 above. In addition to the raw annualized returns, we compute the risk-adjusted alphas using the Fama-French 5 factors and Fama-French 3 factors. We also consider two additional cases of moderate proportional transaction costs during each trade of 0.02% and 0.07%. These figures are motivated from the investigation of Edelen et al. (2013) on the bid-ask spread of liquid US stocks. On top of the reported results, we also carry out various robustness checks (different training samples, quintiles), which leave the results qualitatively unchanged.

[4]This is consistent with the predictive regressions as depicted in Table 8 where the coefficients of $Skew$ have negative signs, while those of IV and Put have positive signs.

Table 10 exhibits the annualized returns of the trading strategies for the case of zero transaction costs. The results are very favorable. For all OCs, the residual-based strategy earns a better Sharpe ratio. For the *Skew*-based residual strategy, we find an annualized Sharpe ratio of 3.2 (versus 2.9), for *IV* 2.6 (versus 1.2), for *Put* 1.5 (versus 1.2). Thus, OCs have both a public and a private information component, whereby the latter can be isolated by regressing the OCs on public information given by market factors and textual sentiment. The Fama-French adjusted returns (alpha) underline furthermore that these results are not driven by common market factors.

When we consider transaction costs of 0.02%, the residual-based strategies still dominate with Sharpe ratios of 2.4, 2.2, and 1.1 (*Skew*, *IV*, *Put*) against 2.3, 1.1, and 1.0, but come off as losers in two out of three cases after incurring transaction costs of 0.07%: 0.8, 1.4, and 0.3 (*Skew*, *IV*, *Put*) against 0.9, 0.8, and 0.6 (tables are omitted for the sake of space). The residual-based strategies gradually lose ground against the purely OC-based ones because residuals vary much more within their rankings than do OCs. Hence, much higher portfolio turnover rates are required and profits dissipate.

In summary, our results suggest that after public information and textual sentiment are filtered from OCs, their unexplained component is highly informative about future stock returns. Thus, we can attach to this isolated private information in option data a significant economic value besides the purely statistical regression evidence. In a practical trading situation, however, it may be eventually difficult to profit from this because of transaction costs.

6 Return predictability and market disagreement

6.1 The market disagreement risk premium

The sentiment index constructed from the firm-level sentiment can be seen as a representative of the average mood in the cross-section. The measurements of firm-level sentiment, however, also convey an additional piece of information: the dispersion of sentiment in the cross-section.

Asset valuation may vary depending on whether the firm-level sentiment is highly concentrated or rather widely dispersed in the cross-section.

From a theoretical point of view, the prediction of how investor disagreement relates to asset returns is controversial.[5] On the one hand, a stream of literature suggests that investors should be compensated for bearing risk if there is disagreement; this could be due to adverse selection and investor heterogeneity (Varian, 1985; David, 2008; Cujean and Hasler, 2016, among others). On the other hand, disagreement in markets could be also be related to lower expected returns. As first articulated by Miller (1977), if pessimists stay out of the market because of short-sale constraints, asset prices reflect only the optimists' valuations and hence are overvalued.

In empirical work, it is common to measure ex-ante disagreement as the standard deviation of analyst forecasts of a particular economic variable of interest, such as future earnings; see, e.g., Park (2005). We follow this approach and compute market disagreement, denoted by $\sigma_{B,t}$, as the standard deviation of the cross-sectional $B_{i,t}$. It is important, however, that our measure of disagreement differ from this approach in that we measure disagreement not in terms of a forecast divergence, but in terms of sentiment heterogeneity: A high value of our disagreement measure on a particular day means that the sentimental firm-level prospects, which are revealed by the articles, are heterogeneous in the cross-section.

In Figure 4, we display some density estimates of trading-hour disagreement; their evolution gives rise to our second-order moment estimates of cross-sectional sentiment. Their correlation with market volatility is remarkably low: -2.6% with SM (-1.0% with LM); hence it is close to orthogonal to market (return) volatility and therefore measures a very different dimension of market uncertainty than does return volatility. The correlation between market disagreement and the sentiment index varies strongly with the approach to extracting sentiment: it is about $+64\%$ for SM, but only $+5\%$ for LM. This is similar to Kim et al. (2014), where disagreement is measured on the basis of the divergence of analyst forecasts.

As discussed above, the literature offers various explanations as to why investor disagreement

[5] See Carlin et al. (2014) for a recent account of the literature.

may impact future asset returns. Here, we take an empirical stance. While a firm sentiment may manifest a signal about a specific firm, heterogeneity in cross-sectional firm-level sentiment implies a source of uncertainty, extracted from news tones, for the market as a whole. We therefore propose

Hypothesis 5 (H5): Cross-sectional sentiment disagreement commands a risk premium.

Using our measurement of disagreement obtained from trading-hour sentiment, we revisit the predictive regressions in Table 11.[6] The regressions contain the same set of control variables, except that we exclude the trading sentiment index $B_{idx,t}$ because the latter is insignificant in most of the predictive regressions. In the regressions, all results stay as reported previously; on top of this, we find that σ_B carries a positive and highly significant coefficient for both the SM and LM case. This lends support to the idea that high levels of disagreement in the cross-sectional distribution of sentiments make investors reluctant to hold assets; hence, similarly to market volatility, they require a positive risk premium if dispersion is high. Because it is almost uncorrelated with market volatility, however, sentiment dispersion is yet another dimension to market uncertainty, here mainly from news tones.

To summarize, our results obtained for sentimental disagreement strongly point in the same direction as those in as Carlin et al. (2014), Cujean and Hasler (2016), among others, who find that disagreement induced by forecast heterogeneity is a positively priced factor.

6.2 Momentum and reversal effects conditional on disagreement states

As pointed out by Cujean and Hasler (2016), disagreement may also generate certain types of return predictability and time series momentum in the sense of Moskowitz et al. (2012). In their model, for instance, large levels of disagreement may cause a short-term momentum effect. Hong

[6] As regards overnight dispersion sentiment, we find qualitatively the same evidence for H5 and H6 of this section. These results are therefore omitted.

and Stein (2007) outline the mechanisms of generating investor disagreement. Gradual information flow, limited attention and heterogeneous prior beliefs are the possible causes. Disagreement therefore becomes decisive for the momentum or reversal effects. To investigate the future price movement conditional on disagreement status, we formulate the following hypothesis:

Hypothesis 6 (H6): Extreme levels of market disagreement evidence return predictability

In order to distinguish between periods of low and high disagreement, we construct two thresholds σ_{10} and σ_{90} as the 10% and 90% percentiles of $\sigma_{B,t}$ and interact them with single-stock returns. The interaction of the two dummy variables is given by

$$\begin{aligned} \sigma^{+}_{10,i,t} &= \mathbf{I}(\sigma_t < \sigma_{10})\, \mathbf{I}(R_{i,t} \geq 0) \\ \sigma^{+}_{90,i,t} &= \mathbf{I}(\sigma_t > \sigma_{90})\, \mathbf{I}(R_{i,t} \geq 0)\ . \end{aligned} \tag{6}$$

We define $\sigma^{-}_{10,i,t}$ and $\sigma^{-}_{90,i,t}$ analogously; $\sigma^{+}_{10,i,t}$ and $\sigma^{-}_{10,i,t}$ stand for a very low disagreement state, i.e., consensus, given the current positive/negative return, whereas $\sigma^{+}_{90,i,t}$ and $\sigma^{-}_{90,i,t}$ represent a high disagreement interacted with current positive/negative return. The interaction between disagreement levels and returns allows for an investigation of momentum and reversal effects conditional on certain levels of disagreement.

In Table 12, we report the results of interacting past returns with extreme levels of market disagreement. First, we check in scenarios (1) and (2) whether the reported sign on disagreement continues to hold in the high disagreement scenarios. We find significant negative signs on $\mathbf{I}_{\sigma_{B_i}<\sigma_{10}}$ and a positive sign on $\mathbf{I}_{\sigma_{B_i}>\sigma_{90}}$, which is in line with the previous results. In scenarios (3) and (4), the interaction terms emerge as significant predictors in times when sentiment is very concentrated (low dispersion or high consensus). More specifically, we find a strong momentum effect for negative returns and a reversal effect for positive returns, respectively, when sentiment is concentrated: if dispersion is low and a negative return is observed, stocks are more likely to experience price continuation (scenario (3)). The same holds if dispersion is low and a positive return is observed (scenario (4)). On the other hand, when dispersion is high, we find a price

reversal on negative returns, while the estimated coefficient on positive returns is insignificant. Squaring these findings with the estimates from LM-based sentiment in scenarios (7) to (12), we find corroborative evidence whenever sentiment is concentrated. The results, however, disagree in heterogeneous sentiment environments.

In summary, whereas our predictive regressions have revealed no evidence of time series momentum (cf. Tables 8 and 9), we find conclusive evidence of return predictability conditional on times of concentrated sentiment. In particular, in low disagreement states, momentum and reversal effects are not exclusive of each other. The interaction with the return makes the two effects distinguishable and contributes to a better comprehension of the informational content of sentiment disagreement. The results in a heterogeneous sentiment environment are not clear-cut and need to be read with caution.

7 Conclusion

The informational content of option characteristics (OCs) and their predictive power for stock returns have often been attributed to the alleged content of private information. Yet option data also embed public information and sentiment. In order to isolate the private information ingrained in option data, we control for publicly available news and their textual sentiment. By this design, we are able to build up a series of testable hypotheses about the role of sentiment and private information in single-stock option markets and equity markets.

To extract public sentiment, we apply supervised and unsupervised learning algorithms to a rich source of NASDAQ articles referring to 97 S&P100 firms. Studying the predictive power of firm-level sentiment, aggregate sentiment, and sentimental disagreement as well as that of classical OCs, such as implied volatility, out-of-the-money put options, and the implied volatility skew, we find that single-stock OCs react to both firm-level and aggregate sentiment. In predictive return regressions, we show that aggregate sentiment indices and OCs predict stock returns jointly and that there is a substantial inside information component in OCs that cannot be accounted for

by public information and sentiment. We add support to these conclusions by showing that a trading strategy based on option information, where sentiment and public information is partialed out, yields higher Sharpe ratios than the standard strategy based on OCs only. We also demonstrate that sentiment disagreement commands a positive risk premium, above and beyond market risk. This lends support to the idea that sentimental disagreement is a risk factor investors ask compensation for and that price momentum or return reversal occurs conditionally on states of concentrated disagreement.

We also shed new light on trading-time versus overnight information. In all predictive regressions, overnight news is more informative than trading-time news. To understand this, we apply a statistical topic model. It shows that while both overnight and trading-time news archives share common topics, they also differ in certain dimensions: overnight articles cover investor information that is of a more fundamental nature, whereas trading-time news additionally includes tactical aspects of the investment process, and both news archives attach different orders of importance to the shared topics. These facts may explain their specific explanatory power.

Overall, first and second-order moments of textual sentiment as uncovered by the analysis of massive text corpora appear to be influential factors for price formation in option markets and equity markets but disseminate at different diffusive speeds. Further research might concentrate on different news or sentiment sources, such as Twitter and StockTwits. A paramount input to news sentiment distillation is of course the underlying lexicon or phrase data bank. For very different asset classes, such as commodities or crypto-currencies, one might therefore need to think of another lexical basis.

In this paper, we have opened a research path towards studying stock return predictability that incorporates machine learning-based sentiments that are distilled from different lexica. Depending on the characteristics of the news they are derived from, such as posting time, topics and topic complexity, these sentiment variables have predictive power and their cross-sectional dispersion commands a strong positive risk premium.

References

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

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

Banerjee, S. (2011). Learning from prices and the dispersion in beliefs, *The Review of Financial Studies* **24**(9): 3025–3068.

Barclay, M. J. and Hendershott, T. (2003). Price discovery and trading after hours, *The Review of Financial Studies* **16**(4): 1041–1073.

Berkman, H. and Truong, C. (2009). Event day 0? after-hours earnings announcements, *Journal of Accounting Research* **47**(1): 71–103.

Bird, S., Klein, E. and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*, " O'Reilly Media, Inc.".

Blei, D. M. (2012). Probabilistic topic models, *Communications of the ACM* **55**(4): 77–84.

Bollerslev, T., Osterrieder, D., Sizova, N. and Tauchen, G. (2013). Risk and return: Long-run relations, fractional cointegration, and return predictability, *Journal of Financial Economics* **108**(2): 409–424.

Bommes, E., Chen, C. Y.-H. and Härdle, W. (2018). Will stocks react to sector specific sentiment?, *submitted to Quantitative Finance* .

Buncic, D. and Gisler, K. I. (2016). Global equity market volatility spillovers: A broader role for the united states, *International Journal of Forecasting* **32**(4): 1317–1339.

Cao, H. H., Coval, J. D. and Hirshleifer, D. (2002). Sidelined investors, trading-generated news, and security returns, *The Review of Financial Studies* **15**(2): 615–648.

Carlin, B. I., Longstaff, F. A. and Matoba, K. (2014). Disagreement and asset prices, *Journal of Financial Economics* **114**(2): 226–238.

Chakravarty, S., Gulen, H. and Mayhew, S. (2004). Informed trading in stock and option markets, *The Journal of Finance* **59**(3): 1235–1257.

Chang, B. Y., Christoffersen, P. and Jacobs, K. (2013). Market skewness risk and the cross section of stock returns, *Journal of Financial Economics* **107**(1): 46–68.

Chen, H., De, P., Hu, Y. J. and Hwang, B.-H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media, *Review of Financial Studies* **27**(5): 1367–1403.

Chen, Y., Chiang, T. and Härdle, W. (2018). Downside risk and stock returns in the G7 countries: an empirical analysis of their long-run and short-run dynamics, *Journal of Banking and Finance* **0**(0): 1–26. Forthcoming.

Conrad, J., Dittmar, R. F. and Ghysels, E. (2013). Ex ante skewness and expected stock returns, *The Journal of Finance* **68**(1): 85–124.

Cujean, J. and Hasler, M. (2016). Why does return predictability concentrate in bad times?, *The Journal of Finance* .

Da, Z., Engelberg, J. and Gao, P. (2014). The sum of all fears investor sentiment and asset prices, *The Review of Financial Studies* **28**(1): 1–32.

Das, S. R. and Chen, M. Y. (2007). Yahoo! for amazon: Sentiment extraction from small talk on the web, *Management science* **53**(9): 1375–1388.

David, A. (2008). Heterogeneous beliefs, speculation, and the equity premium, *The Journal of Finance* **63**(1): 41–83.

Dennis, P. and Mayhew, S. (2002). Risk-neutral skewness: Evidence from stock options, *Journal of Financial and Quantitative Analysis* **37**: 471–493.

Diamond, D. W. and Verrecchia, R. E. (1987). Constraints on short-selling and asset price adjustment to private information, *Journal of Financial Economics* **18**(2): 277–311.

Doyle, J. T. and Magilke, M. J. (2009). The timing of earnings announcements: An examination of the strategic disclosure hypothesis, *The Accounting Review* **84**(1): 157–182.

Edelen, R., Evans, R. and Kadlec, G. (2013). Shedding light on "invisible" costs: Trading costs and mutual fund performance, *Financial Analysts Journal* **69**(1): 33–44.

Engelberg, J. E., Reed, A. V. and Ringgenberg, M. C. (2012). How are shorts informed?: Short sellers, news, and information processing, *Journal of Financial Economics* **105**(2): 260–278.

Han, B. (2008). Investor sentiment and option prices, *Review of Financial Studies* **21**: 387–414.

Hong, H. and Stein, J. C. (2007). Disagreement and the stock market, *Journal of Economic perspectives* **21**(2): 109–128.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews, *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 168–177.

Kim, J. S., Ryu, D. and Seo, S. W. (2014). Investor sentiment and return predictability of disagreement, *Journal of Banking & Finance* **42**: 166–178.

Linton, M., Teo, E. G. S., Bommes, E., Chen, C. and Härdle, W. K. (2017). Dynamic topic modelling for cryptocurrency community forums, *Applied Quantitative Finance*, Springer, pp. 355–372.

Liu, B. (2012). Sentiment analysis and opinion mining, *Synthesis lectures on human language technologies* **5**(1): 1–167.

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.

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

Miller, E. M. (1977). Risk, uncertainty, and divergence of opinion, *The Journal of finance* **32**(4): 1151–1168.

Moshirian, F., Nguyen, H. G. L. and Pham, P. K. (2012). Overnight public information, order placement, and price discovery during the pre-opening period, *Journal of Banking & Finance* **36**(10): 2837–2851.

Moskowitz, T. J., Ooi, Y. H. and Pedersen, L. H. (2012). Time series momentum, *Journal of financial economics* **104**(2): 228–250.

Pan, J. and Poteshman, A. M. (2006). The information in option volume for future stock prices, *The Review of Financial Studies* **19**(3): 871–908.

Park, C. (2005). Stock return predictability and the dispersion in earnings forecasts, *The Journal of Business* **78**(6): 2351–2376.

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.

Sibley, S. E., Wang, Y., Xing, Y. and Zhang, X. (2016). The information content of the sentiment index, *Journal of Banking & Finance* **62**: 164–179.

Stambaugh, R. F., Yu, J. and Yuan, Y. (2012). The short of it: Investor sentiment and anomalies, *Journal of Financial Economics* **104**(2): 288–302.

Stilger, P. S., Kostakis, A. and Poon, S.-H. (2016). What does risk-neutral skewness tell us about future stock returns?, *Management Science* **63**(6): 1814–1834.

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

Tetlock, P. C. (2010). Does public financial news resolve asymmetric information?, *Review of Financial Studies* **23**: 3520–3557.

Varian, H. R. (1985). Divergence of opinion in complete markets: A note, *The Journal of Finance* **40**(1): 309–317.

Wang, X., Wu, C. and Xu, W. (2015). Volatility forecasting: The role of lunch-break returns, overnight returns, trading volume and leverage effects, *International Journal of Forecasting* **31**(3): 609–619.

Wiebe, J. and Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts, *CICLing*, Vol. 5, Springer, pp. 486–497.

Wilson, T., Wiebe, J. and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis, *Proceedings of the conference on human language technology and empirical methods in natural language processing*, Association for Computational Linguistics, pp. 347–354.

Xing, Y., Zhang, X. and Zhao, R. (2010). What does the individual option volatility smirk tell us about future equity returns?, *Journal of Financial and Quantitative Analysis* **45**(3): 641–662.

Yu, J. (2011). Disagreement and return predictability of stock portfolios, *Journal of Financial Economics* **99**(1): 162–183.

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 and Economic Statistics* **34**(4): 547–563.

A Appendix

A.1 List of the 97 companies included in the analysis

Apple Inc. (AAPL); AbbVie Inc. (ABBV); Accenture PLC. (ACN); Automatic Data Processing Inc. (ADP); Aetna Inc. (AET); American International Group Inc. (AIG); Amgen Inc. (AMGN); American Tower Corp. (AMT); Amazon.com (AMZN); Anadarko Petroleum Corp. (APC); American Express Inc. (AXP); Boeing Co. (BA); Bank of America Corp. (BAC); Best Buy Co. Inc. (BBY); Baker Hughes Inc. (BHI); Biogen Inc. (BIIB); Bristol-Myers Squibb (BMY); Citigroup Inc. (C); Caterpillar Inc. (CAT); CBS Corp. (CBS); Celgene Corp. (CELG); Chesapeake Energy Corp. (CHK); Comcast Corp. (CMCSA); Chipotle Mexican Grill Inc. (CMG); ConocoPhillips Co. (COP); Costco Wholesale Corp. (COST); Cisco Systems Inc. (CSCO); CVS Health Corp. (CVS); Chevron (CVX); Delta Air Lines Inc. (DAL); DuPont Inc. (DD); Danaher Corp. (DHR); The Walt Disney Company (DIS); Dow Chemical (DOW); Duke Energy Corp. (DUK); Electronic Arts Inc. (EA); eBay Inc. (EBAY); E-TRADE Financial Corp. (ETFC); Exelon (EXC); Ford Motor (F); FedEx (FDX); First Solar Inc. (FSLR); General Dynamics Corp. (GD); General Electric Co. (GE); Gilead Sciences (GILD); General Motors (GM); Gap Inc. (GPS); Goldman Sachs (GS); Halliburton (HAL); Home Depot (HD); Honeywell (HON); Hewlett-Packard Co. (HPQ); International Business Machines (IBM); Intel Corporation (INTC); Johnson & Johnson Inc. (JNJ); JP Morgan Chase & Co. (JPM); The Coca-Cola Co. (KO); The Kroger Co. (KR); Lennar Corp. (LEN); Eli Lilly (LLY); Lockheed-Martin (LMT); Southwest Airlines Co. (LUV); Macy's Inc. (M); Mastercard Inc. (MA); McDonald's Corp. (MCD); Medtronic Inc. (MDT); 3M Company (MMM); Altria Group Inc. (MO); Merck & Co. (MRK); Morgan Stanley (MS); Microsoft (MSFT); Micron Technology Inc. (MU); Newmont Mining Corp. (NEM); Netflix Inc. (NFLX); NextEra Energy (NKE); Northrop Grumman Corp. (NOC); NVIDIA Corp. (NVDA); Pepsico Inc. (PEP); Pfizer Inc. (PFE); Procter & Gamble Co. (PG); Phillip Morris International (PM); Qualcomm Inc. (QCOM); Starbucks Corp. (SBUX); Schlumberger (SLB); Simon Property Group, Inc. (SPG); AT&T Inc. (T); Target Corp. (TGT); Travelers Cos. Inc. (TRV); Time Warner Inc. (TWX); UnitedHealth Group Inc. (UNH); United Technologies Corp. (UTX); Visa

Inc. (V); Verizon Communications Inc. (VZ); Wells Fargo (WFC); Wal-Mart (WMT); Exxon Mobil Corp. (XOM); Yahoo! Inc. (YHOO).

A.2 Methodological details on sentiment estimation

A.2.1 Lexicon method (LM)

Here, we illustrate the "bag-of-words" approach for a positive sentiment Pos; the calculation is analogous for the negative sentiment Neg. To simplify the presentation, assume that the textual data only contain articles regarding the subject of interest, e.g., a specific company i. Consider a collection of texts $D_{i,t}$ with $j = 1, \ldots, J$ unique words W_j about i. The number of appearances of W_j at t for i, denoted by $w_{i,t,j}$, is counted and the total number of words for company i on day t is calculated as $N_{i,t} = \sum_{j=1}^{J} w_{i,t,j}$. Then one proceeds to measure the positive sentiment using the fraction of positive words per day:

$$Pos_{i,t} = N_{i,t}^{-1} \sum_{j=1}^{J} \mathbf{I}\left(W_j \in L_{Pos}\right) w_{i,t,j} \ , \tag{7}$$

where L_{Pos} denotes the set of positive words in a predefined dictionary. Dictionaries that are widely used are described, e.g., in Loughran and McDonald (2011), Liu (2012), or Zhang et al. (2016).

Eq. (7) is usually adjusted to account for negation, as for example the term `not good` lacks a positive meaning. In practice, negation is often handled by looking at the n-gram, a sequence of n words around a lexical element $W_j \in L$, with L a lexicon. One can see that the position in the text matters for such an approach and words may not be re-ordered until negated words in L are counted. Thus, if the distance between a sentiment word and a negation word is less than a prespecified threshold, the polarity of the word is inverted as suggested, e.g., in Hu and Liu (2004). We give a concrete example below in Section 2.2.

Specifically, if L_{Neg} and L_{Pos} are the sets of negative and positive words, respectively, and addi-

tionally, $f_{i,t,j}$ and $u_{i,t,j}$ account, respectively, for the frequency of negated negative and negated positive words in $D_{i,t}$ we refine (7) as:

$$Pos_{i,t} = N_{i,t}^{-1} \sum_{j=1}^{J} \left\{ \mathbf{I}\left(W_j \in L_{Pos}\right) \left(w_{i,t,j} - u_{i,t,j}\right) \; + \; \mathbf{I}\left(W_j \in L_{Neg}\right) f_{i,t,j} \right\}, \tag{8}$$

in which negated negative words are treated as positive and negated positive words as negative.

As explained in the main text, a sentence level is more precise (Wiebe and Riloff, 2005; Wilson et al., 2005). We therefore switch the focus from a word-based to a sentence-based polarity. More precisely, fix a company i and a date t, drop these indices for notational simplicity, and define (in abuse of the index j) as in (7) and (8) the positive/negative sentiment on the sentence level of a given document. Then calculate for each sentence j, $j = 1, \ldots, n$, its polarity as

$$Pol_j = \mathbf{I}(Pos_j > Neg_j) - \mathbf{I}(Pos_j < Neg_j)$$

and finally aggregate as

$$FP = n^{-1} \sum_{j=1}^{n} \mathbf{I}(Pol_j = 1) \tag{9}$$

$$FN = n^{-1} \sum_{j=1}^{n} \mathbf{I}(Pol_j = -1), \tag{10}$$

where n is the number of sentences in the document. Eqs. (9) and (10) indicate the fraction of positive (FP) and negative (FN) polarity of company i at date t, which is used to compute Eq. (1).

A.2.2 Supervised method (SM)

The basis of the supervised learning approach is the financial phrase bank of Malo et al. (2014). The basic difference from the LM method is that polarity (denoted here as Y) is given through a set of annotators that fix the sentiment of sentences like `The profit of Apple increased` or `The profit of the company decreased` in an experimental set-up.

Let us explain the numerisization of these sentences in more detail. We first lemmatize the words and employ 1-grams and 2-grams to create the word vector $X = ($ `the, profit, of, apple, increased, company, decreased, the profit, profit of, of the, of apple, the company, apple increase, company decrease` $)^\top$ in 14 dimensions. The two sentences above then result in the vectors

$$X_1 = (1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, 1, 0)^\top$$
$$\text{and} \qquad X_2 = (2, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1)^\top .$$

These sentences have the obvious (but human-annotated) outcome $Y_1 = 1$ for X_1 and $Y_2 = -1$ for X_2. We thus can define a score-based discrete response model. The score for a parameter vector β is $s(X) = \beta^\top X$, $\beta \in \mathbb{R}^p$ with a possibly large dimension p.

Following Luhn (1957), the word matrix consisting of all sentences is then transformed into a $tf - idf$ matrix. Since the sentiment may be either negative, neutral or positive, we have to run the predictive model involving $s(X)$ three times. More precisely, we put $Y = 1$ for positive and $Y = -1$ for both neutral and negative. Then we put $Y = 1$ for neutral sentiment and $Y = -1$ for the rest. Finally, $Y = 1$ for negative sentiments and $Y = -1$ for the remaining positive and neutral sentiments. Each of the three resulting scores will give us a probability of misclassification or a confidence score. We finally pick the score with the best confidence.

To be more specific about estimation, given a regularized linear model, the training data (X_1, Y_1), $\ldots, (X_n, Y_n)$ with $X_i \in \mathbb{R}^p$ and $Y_i \in \{-1, 1\}$, and the linear scoring function $s(X)$, we calibrate the predictive model via the regularized training error

$$n^{-1} \sum_{i=1}^{n} L\{Y_i, s(X)\} + \lambda R(\beta) \tag{11}$$

with $L(\cdot)$ as loss function, $R(\cdot)$ as regularization term and penalty $\lambda \geq 0$. We have applied different loss functions. In terms of support vector machines (SVM), one may employ the Hinge

loss

$$L\{Y, s(X)\} = \max\{0,\, 1 - s(X)Y\} \tag{12}$$

or the Logistic likelihood $L(u) = \exp(u)/\{1 + \exp(u)\}$. The least squares loss $L(u) = u^2$ leads to the well known ridge regression. As a regularization term one may employ the L_2 norm $R(\beta) = p^{-1} \sum_{i=1}^{p} \beta_i^2$ or the L_1 norm $R(\beta) = \sum_{i=1}^{p} |\beta_i|$, giving the calibration task a Lasso type twist.

The question now arises of how to determine the loss functions L, the regularization term R and the hyper parameter λ. We calibrated (11) for the described set of L, R functions using the Stochastic Gradient Descent (SGD) method. The Malo et al. (2014) training data set is available at https://www.researchgate.net/publication/251231364_FinancialPhraseBank-v10 and the Python code is described on http://www.quantlet.de in `TXTfpbsupervised`. The regularization parameter was optimized using 5-fold cross-validation in which the data set is partitioned into 5 complementary subsets. Four out of these 5 subsets were then combined to build the training data set. Furthermore, we oversampled sentences with positive and negative sentiment in the training set to obtain a balanced sample and control for the trade off between the type 1 and type 2 error. In summary, we ran 66K predictive models and obtained the best supervised learning accuracy for the hinge loss and the L_1 penalty.

Table 1: **Confusion matrices of the SM and LM methods**

True \ Pred	SM with Oversampling				LM			
	−1	0	1	Total	−1	0	1	Total
−1	1992	289	254	2535	213	289	12	514
0	96	2134	305	2535	200	2187	148	2535
1	105	469	1961	2535	111	772	285	1168
Total	2184	2901	2520	7605	524	3248	445	4217
Precision	0.91	0.74	0.78		0.41	0.67	0.64	
Recall	0.78	0.84	0.77		0.41	0.86	0.24	

Negative sentences are oversampled in order to yield a comparable number of negative sentences as there are positive ones in the Malo et al. (2014) training data set. A 5-fold cross validation is employed to avoid overfitting. The best model is the one with the highest precision and recall on the manually labeled training data set. Precision is defined as the ratio of true positives to the sum of true positives and false positives, which is equivalent to 1−type 2 error. Recall is a ratio of true positives to the sum of true positives and false negatives, equivalent to 1−type 1 error.

Table 2: **Topic Model Fit to Overnight Articles**

	Topics and most frequent words									
	1	2	3	4	5	6	7	8	9	10
Topics	*Dividends*	*Inv. stratg.*	*Earnings*	*Equities*	*Asset mgmt*	*Econ. Outlook*	*Charts*	*Anal. Roundup*	*Sectors*	*Market*
	dividend	**stock**	**earnings**	tale	**fund**	**stocks**	**average**	**analyst**	**update**	**market**
	ex	reasons	**estimates**	tape	**income**	buy	**moving**	**blog**	**sector**	**report**
	date	**focus**	follow	**continue**	**municipal**	**oil**	day	**growth**	**energy**	pre
	scheduled	**great**	**history**	**higher**	**nuveen**	higher	**cross**	new	**health**	**nasdaq**
	corporation	**investors**	**indicator**	**shares**	**dividend**	**week**	**bullish**	data	**care**	**index**
	september	choice	reaction	**focus**	**ex**	best	notable	**beat**	**financial**	**close**
	june	**value**	**sensitive**	**estimates**	**scheduled**	**news**	makes	**shares**	**consumer**	active
Top 15 words	**march**	**jumps**	**revenues**	march	**date**	**data**	**critical**	**energy**	ung	**composite**
	november	session	beat	**surge**	high	lower	**breaks**	**high**	uso	**closes**
	august	**growth**	beats	**strong**	new	**ahead**	key	week	**technology**	**points**
	trust	**momentum**	season	**value**	**eaton**	**watch**	**level**	miss	**close**	qqq
	february	rises	**surprise**	**great**	**vance**	today	crosses	loss	closing	**aapl**
	december	right	revenue	**growth**	**trust**	**china**	alert	**roundup**	**oil**	**bac**
	july	adds	strong	**falls**	quality	**dividend**	**crossover**	**revenues**	partners	xiv
	october	moves	misses	**holdings**	ii	**growth**	dow	estimates	dis	tvix

Results of the topic model fit (Latent Dirichlet Allocation) to overnight articles. The columns feature the 10 topics in order of frequency. Each column displays the 15 most important words of the respective topic, again in order of frequency. Italicized topic labels are based on our interpretation of the empirical word sets.

Table 3: **Topic Model Fit to Trading Time Articles**

	Topics and most frequent words									
	1	2	3	4	5	6	7	8	9	10
Topics	*Press rel.*	*Earnings 1*	*Funds*	*Option trades*	*Charts*	*Sectors*	*Dividends*	*Equities*	*Earnings 2*	*Share types*
	analyst	**earnings**	**etf**	**options**	**average**	update	**stock**	**stocks**	**indicator**	**shares**
	blog	**revenues**	detected	**trading**	**moving**	**sector**	**reminder**	**buy**	**earnings**	cross
	zacks	**beat**	big	**using**	day	**energy**	market	**new**	follow	**yield**
	highlights	**estimates**	**inflow**	week	cross	**financial**	**preferred**	strong	**history**	**series**
	releases	**beats**	**inflows**	**interesting**	**bullish**	**technology**	today	**oil**	**reaction**	mark
	press	**miss**	**outflow**	**earn**	notable	**consumer**	**series**	mid	**sensitive**	**preferred**
	energy	season	**outflows**	commit	**critical**	**health**	news	**sell**	corp	dma
Top 15 words	**group**	**report**	notable	**buy**	makes	**care**	**ex**	**etfs**	**corporation**	**dividend**
	holdings	view	large	annualized	**breaks**	mid	**cumulative**	european	company	today
	international	store	noteworthy	**available**	key	**market**	**dividend**	**adrs**	international	mid
	high	**sales**	alert	begin	**crosses**	afternoon	interesting	day	**group**	**cumulative**
	american	misses	experiences	**purchase**	level	day	corp	news	systems	**ex**
	loss	tops	**ishares**	october	**crossover**	laggards	**roundup**	**market**	technology	higher
	week	**surprise**	**etfs**	january	alert	**oil**	**redeemable**	gains	**holdings**	afternoon
	airlines	**revenue**	**spdr**	november	option	morning	non	higher	technologies	reminder

Results of the topic model fit (Latent Dirichlet Allocation) to trading time articles. The columns feature the 10 topics in order of frequency. Each column displays the 15 most important words of the respective topic, again in order of frequency. Italicized topic labels are based on our interpretation of the empirical word sets.

Table 4: **Descriptive Statistics**

	Summary Statistics					
	Variable	Mean	25%	50%	75%	Std
Supervised learning	B_i	11.26	0.00	0.00	23.08	18.32
	BN_i	0.63	0.00	0.00	0.00	4.15
	B_{idx}	11.26	8.82	11.26	13.57	3.39
	BN_{idx}	0.63	0.12	0.44	0.90	0.65
	B_i^{on}	10.88	0.00	0.00	22.24	16.81
	BN_i^{on}	0.39	0.00	0.00	0.00	3.03
	B_{idx}^{on}	10.88	9.06	10.80	12.62	2.87
	BN_{idx}^{on}	0.39	0.09	0.30	0.60	0.38
Lexicon projection	B_i	1.12	0.00	0.00	0.80	15.52
	BN_i	3.46	0.00	0.00	0.00	9.81
	B_{idx}	1.12	−0.57	1.08	2.77	2.44
	BN_{idx}	3.46	2.21	3.39	4.43	1.67
	B_i^{on}	3.42	0.00	0.00	6.17	12.99
	BN_i^{on}	1.83	0.00	0.00	0.00	6.71
	B_{idx}^{on}	3.42	1.65	3.34	5.10	2.54
	BN_{idx}^{on}	1.83	1.12	1.69	2.38	0.95
OC	$Skew$	5.83	3.81	5.45	7.40	3.33
	Put	0.57	0.19	0.35	0.67	0.73
	IV	24.07	17.03	21.39	28.19	10.49

Descriptive statistics of sentiment variables for both the supervised learning and the lexicon projection method and option characteristics (OC) during the sample period Jan. 2012 to Apr. 2016, all expressed in %-terms. B_i is daily bullishness, BN_i negative daily bullishness, while B_{idx} and BN_{idx} denote the respective bullishness indices over all 97 firms. Superscript *on* distinguishes overnight measures from trading time measures. IV is implied volatility, $Skew$ the implied volatility skew, and Put the relative put price as defined in the main text. Source: NASDAQ articles, IvyMetrics US (OptionMetrics), own computations.

Table 5: **Correlation between OC and sentiment variables**

		$Skew_{i,t}$		$IV_{i,t}$		$Put_{i,t}$	
		SM	LM	SM	LM	SM	LM
$B_{i,t}$	25%	−4.24	−2.80	−6.30	−9.19	−5.01	−6.14
	50%	−1.28	0.20	−1.99	−4.66	−2.05	−3.77
	75%	1.91	3.22	2.44	1.21	1.24	0.00
$B_{idx,t}$	25%	−6.74	−3.63	−19.81	−14.40	−14.02	−9.98
	50%	−2.82	−0.24	−9.68	−9.27	−5.68	−6.52
	75%	4.61	3.34	4.10	−2.92	3.49	−2.22
$BN_{idx,t}$	25%	−6.49	−9.17	7.01	0.39	9.13	4.56
	50%	−2.87	−3.12	17.57	9.25	16.04	10.21
	75%	3.38	3.07	32.32	24.88	28.08	19.72
$B^{on}_{i,t}$	25%	−3.17	−2.66	−3.82	−9.19	−3.56	−7.79
	50%	−0.30	0.25	−0.32	−4.66	−0.50	−3.70
	75%	2.26	3.15	3.61	1.21	3.86	0.51
$B^{on}_{idx,t}$	25%	−2.51	−6.40	−12.48	−27.73	−5.62	−20.74
	50%	1.39	−1.12	−2.99	−15.37	0.09	−12.70
	75%	5.61	5.68	6.35	−0.62	5.70	−1.02
$BN^{on}_{idx,t}$	25%	−7.67	−2.87	−5.07	7.51	−5.52	8.03
	50%	−4.52	−0.08	2.10	10.69	−0.03	10.61
	75%	−1.97	2.44	6.88	14.55	4.39	13.01

Correlations of sentiment variables for both the supervised learning and the lexicon projection method and option characteristics (OC) during the sample period Jan. 2012 to Apr. 2016, all expressed in %-terms. B_i is daily bullishness, BN_i negative daily bullishness, while B_{idx} and BN_{idx} denote the respective bullishness indices over all 97 firms. Superscript *on* distinguishes overnight measures from trading time measures. IV is implied volatility, $Skew$ the implied volatility skew, and Put the relative put price as defined in the main text. Source: NASDAQ articles, IvyMetrics US (OptionMetrics), own computations.

Table 6: **OCs and sentiment based on supervised method**

	$Skew_{i,t}$			$Put_{i,t}$			$IV_{i,t}$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$B_{i,t}$	−0.0186	−0.0179	−0.0066	−0.0082	−0.0081	−0.0064	−0.1064	−0.1046	−0.0647
	0.022	0.027	0.452	0.000	0.000	0.000	0.000	0.000	0.012
$BN_{idx,t}$			0.3909			0.3338			4.5073
			0.000			0.000			0.000
$B_{idx,t}$			−0.0759			−0.0228			−0.3986
			0.000			0.000			0.000
MKT		0.0036	0.0044		0.0000	0.0005		−0.0047	0.0028
		0.000	0.000		0.776	0.000		0.000	0.000
SMB		0.0007	0.0006		0.0002	0.0002		0.0020	0.0015
		0.005	0.007		0.000	0.000		0.000	0.000
HML		0.0009	0.0011		0.0000	0.0002		0.0000	0.0027
		0.003	0.000		0.807	0.004		0.995	0.726
RMW		−0.0001	0.0000		−0.0001	−0.0001		0.0011	0.0002
		0.854	0.928		0.319	0.000		0.228	0.000
CMA		0.0034	0.0033		0.0007	0.0008		0.0061	0.0065
		0.000	0.000		0.000	0.000		0.000	0.000
R^2 (%)	0.01	0.46	0.50	0.01	0.12	0.66	0.01	0.16	0.77

Sentiment-related variables are quantified by SM. Instrumental variable fixed effects panel regressions with lagged $B_{i,t-1}$, $B_{idx,t-1}$, and $BN_{idx,t-1}$ used as instruments for $B_{i,t}$, $B_{idx,t}$, $BN_{idx,t}$, respectively. All regressions contain a constant and fixed effects. In total, we have 82253 daily observations, and 97 ticker symbols. Below each estimate the p-value based on robust standard errors is displayed.

Table 7: **OCs and sentiment based on lexicon method**

	$Skew_{i,t}$			$Put_{i,t}$			$IV_{i,t}$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$B_{i,t}$	−0.0166	−0.0185	−0.0110	−0.0350	−0.0351	−0.0289	−0.4893	−0.4867	−0.3695
	0.000	0.400	0.684	0.000	0.000	0.000	0.000	0.000	0.000
$BN_{idx,t}$			−0.0806			0.0378			0.3794
			0.057			0.027			0.061
$B_{idx,t}$			−0.0565			−0.0200			−0.4651
			0.150			0.060			0.001
MKT		0.0035	0.0035		0.0000	0.0000		−0.0050	−0.0045
		0.000	0.000		0.680	0.436		0.000	0.000
SMB		0.0007	0.0007		0.0002	0.0002		0.0021	0.0023
		0.004	0.003		0.000	0.000		0.001	0.000
HML		0.0009	0.0009		−0.0001	−0.0001		−0.0008	−0.0019
		0.004	0.005		0.431	0.046		0.346	0.011
RMW		−0.0001	0.0001		0.0000	−0.0001		0.0013	0.0011
		0.781	0.860		0.632	0.217		0.323	0.297
CMA		0.0034	0.0031		0.0006	0.0005		0.0042	0.0025
		0.000	0.000		0.000	0.000		0.011	0.071
R^2 (%)	0.01	0.51	0.60	0.07	0.07	0.15	0.05	0.07	0.19

Sentiment-related variables are quantified by LM. Instrumental variable fixed effects panel regressions with lagged $B_{i,t-1}$, $B_{idx,t-1}$, and $BN_{idx,t-1}$ used as instruments for $B_{i,t}$, $B_{idx,t}$, $BN_{idx,t}$, respectively. All regressions contain a constant and fixed effects. In total, we have 82253 daily observations, and 97 ticker symbols. Below each estimate the p-value based on robust standard errors is displayed.

Table 8: **Predictive regressions with the OCs and sentiment variables**

	$R_{i,t+1}$								
				SM			LM		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$B_{i,t}$				−0.0003	−0.0002	−0.0002	−0.0007	−0.0006	−0.0007
				0.382	0.506	0.499	0.058	0.084	0.073
$BN_{idx,t}$				−0.0686	−0.0708	−0.0694	−0.0237	−0.0244	−0.0232
				0.000	0.000	0.000	0.000	0.000	0.000
$B_{idx,t}$				−0.0014	−0.0013	−0.0010	0.0040	0.0036	0.0044
				0.458	0.508	0.600	0.220	0.275	0.178
$B^{on}_{i,t}$				−0.0005	−0.0003	−0.0003	−0.0004	−0.0003	−0.0003
				0.181	0.371	0.372	0.452	0.532	0.562
$BN^{on}_{idx,t}$				−0.0407	−0.0337	−0.0343	0.0081	0.0075	0.0084
				0.013	0.042	0.038	0.298	0.330	0.279
$B^{on}_{idx,t}$				0.0092	0.0092	0.0095	0.0071	0.0082	0.0084
				0.000	0.000	0.000	0.017	0.007	0.005
$Skew_{i,t}$	−0.0036			−0.0041			−0.0040		
	0.109			0.070			0.076		
$Put_{i,t}$		0.0854			0.0859			0.0872	
		0.004			0.004			0.003	
$IV_{i,t}$			0.0063			0.0063			0.0064
			0.000			0.000			0.000
$R_{i,t}$	0.0123	0.0123	0.0128	0.0125	0.0125	0.0130	0.0126	0.0126	0.0131
	0.241	0.239	0.220	0.231	0.233	0.214	0.229	0.228	0.210
$\log \sigma^2_{i,t}$	0.0006	0.0000	−0.0002	0.0007	0.0000	−0.0002	0.0006	0.0000	−0.0002
	0.001	0.947	0.276	0.001	0.901	0.310	0.001	0.984	0.233
$\log \sigma^2_{mkt,t}$	0.0017	0.0017	0.0018	0.0021	0.0021	0.0022	0.0022	0.0022	0.0023
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
R^2 (%)	0.21	0.30	0.29	0.30	0.40	0.38	0.28	0.38	0.36

Sentiment-related variables appearing in (4) to (6) are quantified by SM, while those in (7) to (9) are projected by LM. All regressions include a global constant, Fama-French 5 factors, but no FE fixed effects (F-test indicates FE are jointly zero). Below each estimate the p-value based on robust standard errors is displayed.

Table 9: **Sources of Predictability**

	$R_{i,t+1}$								
				SM			LM		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$B_{i,t}$				−0.0003	−0.0003	−0.0003	−0.0007	−0.0007	−0.0007
				0.391	0.393	0.378	0.062	0.062	0.062
$BN_{idx,t}$				−0.0685	−0.0683	−0.0680	−0.0235	−0.0238	−0.0237
				0.000	0.000	0.000	0.000	0.000	0.000
$B_{idx,t}$				−0.0013	−0.0014	−0.0012	0.0039	0.0038	0.0039
				0.490	0.485	0.522	0.224	0.237	0.224
$B^{on}_{i,t}$				−0.0005	−0.0005	−0.0005	−0.0004	−0.0004	−0.0004
				0.179	0.188	0.185	0.442	0.440	0.448
$BN^{on}_{idx,t}$				−0.0393	−0.0391	−0.0389	0.0080	0.0075	0.0074
				0.017	0.018	0.018	0.299	0.335	0.337
$B^{on}_{idx,t}$				0.0092	0.0090	0.0090	0.0071	0.0071	0.0072
				0.000	0.000	0.000	0.018	0.017	0.016
$Skew^{\perp}_{i,t}$	−0.0043			−0.0044			−0.0043		
	0.063			0.057			0.064		
$Put^{\perp}_{i,t}$		0.1189			0.1185			0.1192	
		0.001			0.001			0.001	
$IV^{\perp}_{i,t}$			0.0122			0.0121			0.0122
			0.000			0.000			0.000
$R_{i,t}$	0.0119	0.0116	0.0116	0.0122	0.0118	0.0119	0.0122	0.0119	0.0119
	0.253	0.265	0.262	0.245	0.256	0.253	0.242	0.253	0.250
$\log \sigma^2_{i,t}$	0.0007	0.0002	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\log \sigma^2_{mkt,t}$	0.0017	0.0017	0.0017	0.0020	0.0020	0.0020	0.0021	0.0021	0.0021
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
R^2 (%)	0.21	0.34	0.36	0.30	0.43	0.45	0.28	0.41	0.43

$Skew^{\perp}_{i,t}$ is estimated as the residuals by regressing $Skew_{i,t}$ on $\mathbf{B}_t$ and control variables $X_{i,t}$. Likewise, $Put^{\perp}_{i,t}$ and $IV^{\perp}_{i,t}$ can be estimated in the same way. $Skew^{\perp}_{i,t}$, $Put^{\perp}_{i,t}$ and $IV^{\perp}_{i,t}$ are orthogonal to public information and adjusted for the market-wide risk premium. All regressions include a global constant, Fama-French 5 factors, but no FE fixed effects (F-test indicates FE are jointly zero).

Table 10: **Performance of trading strategies**

	Trading strategies					
	Skew residual			*Skew*		
	Long-Short	FF_5	FF_3	Long-Short	FF_5	FF_3
Daily Return (in bp)	14.42	14.74	14.77	14.18	14.61	14.58
P value	0.002	0.002	0.002	0.004	0.004	0.004
Ann. Return	0.43	0.45	0.45	0.43	0.44	0.44
Daily Vol. (in bp)	86.25			92.79		
Ann. Vol.	0.14			0.15		
Daily Sharpe Ratio	0.17			0.15		
Ann. Sharpe Ratio	3.18			2.91		
	IV residual			*IV*		
	Long-Short	FF_5	FF_3	Long-Short	FF_5	FF_3
Daily Return (in bp)	12.41	12.54	12.57	6.79	7.14	7.26
P value	0.009	0.010	0.010	0.181	0.121	0.141
Ann. Return	0.36	0.37	0.37	0.19	0.20	0.20
Daily Vol. (in bp)	88.67			99.28		
Ann. Vol.	0.14			0.16		
Daily Sharpe Ratio	0.14			0.07		
Ann. Sharpe Ratio	2.59			1.18		
	Put residual			*Put*		
	Long-Short	FF_5	FF_3	Long-Short	FF_5	FF_3
Daily Return (in bp)	7.43	7.86	7.70	6.52	6.92	6.87
P value	0.098	0.090	0.098	0.178	0.118	0.140
Ann. Return	0.20	0.22	0.21	0.18	0.19	0.19
Daily Vol. (in bp)	85.66			94.18		
Ann. Vol.	0.14			0.15		
Daily Sharpe Ratio	0.09			0.07		
Ann. Sharpe Ratio	1.51			1.19		

Returns and Sharpe ratios for trading strategies on a daily basis when OC is skew, implied volatility (IV), and the OTM put. Zero transaction costs. "Ann." is short for "Annualized", "Vol." is short for "Volatility", and "bp" is short for "basis points". The daily (annualized) Sharpe ratio is calculated by dividing the daily (annualized) return by the daily (annualized) volatility. Left panel features residual-based strategies, right panel strategies that are based directly on the option characteristic. The columns named "Long-Short" exhibit the figures as calculated on the raw returns of the strategy, while FF_5 and FF_3 means the returns are adjusted by Fama-French 5 factors and Fama-French 3 factors, respectively.

Table 11: **Market consensus and return predictability**

	$R_{i,t+1}$					
		SM			LM	
	(1)	(2)	(3)	(4)	(5)	(6)
$B_{i,t}$	−0.0006	−0.0006	−0.0006	−0.0009	−0.0009	−0.0009
	0.103	0.094	0.092	0.018	0.017	0.016
$BN_{idx,t}$	−0.0814	−0.0825	−0.0819	−0.0505	−0.0515	−0.0520
	0.010	0.000	0.000	0.000	0.000	0.000
$B^{on}_{idx,t}$	0.0071	0.0068	0.0069	0.0032	0.0031	0.0031
	0.001	0.001	0.001	0.253	0.269	0.274
$BN^{on}_{idx,t}$	−0.0445	−0.0446	−0.0442	0.0069	0.0063	0.0061
	0.006	0.006	0.007	0.371	0.418	0.426
σ_{B_i}	0.0112	0.0123	0.0120	0.0177	0.0173	0.0184
	0.000	0.000	0.000	0.000	0.000	0.000
$Skew^{\perp}_{i,t}$	−0.0042			−0.0042		
	0.071			0.072		
$Put^{\perp}_{i,t}$		0.1207			0.1207	
		0.001			0.001	
$IV^{\perp}_{i,t}$			0.0123			0.0124
			0.000			0.000
$R_{i,t}$	0.0122	0.0119	0.0118	0.0122	0.0118	0.0119
	0.245	0.255	0.253	0.245	0.256	0.253
$\log \sigma^2_{i,t}$	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007
	0.000	0.000	0.000	0.000	0.000	0.000
$\log \sigma^2_{mkt,t}$	0.0022	0.0021	0.0022	0.0023	0.0023	0.0023
	0.000	0.000	0.000	0.000	0.000	0.000
R^2 (%)	0.33	0.46	0.48	0.32	0.46	0.48

Predictive stock return regressions. All regressions include a global constant, Fama-French 5 factors, but no FE fixed effects. σ_{B_i} denotes the cross-sectional dispersion of firm-specific sentiment. SM versus LM distinguish sentiment quantified by supervised learning and lexicon projection, respectively. For further annotations; see Table 9. Sample size $N = 82253$ across 97 groups. Below each estimate the p-value based on robust standard errors is displayed.

Table 12: **Market consensus, return and return predictability**

	$R_{i,t+1}$											
	SM						LM					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
$B_{i,t}$	−0.0005	−0.0005	−0.0005	−0.0004	−0.0005	−0.0004	−0.0008	−0.0006	−0.0007	−0.0007	−0.0006	−0.0006
	0.119	0.176	0.165	0.201	0.185	0.258	0.027	0.088	0.054	0.051	0.089	0.080
$BN_{idx,t}$	−0.0761	−0.0777	−0.074	−0.070	−0.079	−0.0721	−0.040	−0.0229	−0.0330	−0.0336	−0.0243	−0.0274
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$B^{on}_{idx,t}$	0.006	0.009	0.0071	0.0075	0.0085	0.008	0.002	0.008	0.005	0.005	0.008	0.008
	0.002	0.000	0.001	0.000	0.000	0.000	0.412	0.003	0.056	0.073	0.006	0.005
$BN^{on}_{idx,t}$	−0.0398	−0.0448	−0.040	−0.041	−0.043	−0.0456	0.0006	0.0084	0.0060	0.0052	0.0082	0.0092
	0.014	0.006	0.015	0.012	0.009	0.005	0.934	0.277	0.444	0.501	0.293	0.237
$\mathbf{I}_{\sigma_{B_i} \leq \sigma_{10}}$	−0.001						−0.002					
	0.000						0.000					
$\mathbf{I}_{\sigma_{B_i} \geq \sigma_{90}}$		0.001						−0.0005				
		0.000						0.025				
$\sigma^{-}_{10,i,t}$			−0.001						−0.0017			
			0.000						0.000			
$\sigma^{+}_{10,i,t}$				−0.001						−0.0020		
				0.017						0.000		
$\sigma^{-}_{90,i,t}$					0.002						−0.0008	
					0.000						0.007	
$\sigma^{+}_{90,i,t}$						0.0000						0.0000
						0.993						0.880
$Skew^{\perp}_{i,t}$	−0.0045	−0.0041	−0.005	−0.005	−0.0043	−0.0046	−0.0041	−0.0043	−0.0042	−0.0044	−0.0045	−0.0045
	0.054	0.076	0.052	0.050	0.065	0.048	0.080	0.061	0.073	0.058	0.054	0.058
R^2 (%)	0.34	0.33	0.33	0.31	0.35	0.31	0.39	0.28	0.31	0.33	0.29	0.28

Predictive stock return regressions. All regressions include a global constant, time-t firm-level returns, idiosyncratic volatility, market volatility, Fama-French 5 factors, but no FE fixed effects. $\mathbf{I}_{\sigma_{B_i}<x}$ denotes a quantile of the cross-sectional dispersion of firm-specific sentiment and σ^{-}_{10} and σ^{+}_{10} are their interactions with firm-specific returns, see (6) for the definition. SM versus LM distinguish sentiment quantified by supervised learning and lexicon projection, respectively. For further annotations, see Table 9. Below each estimate the p-value based on robust standard errors is displayed.

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.

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.