Media-expressed Tone, Option Characteristics, and Stock Return Predictability [*]

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

Abstract

We investigate the informational content of a huge assortment of NASDAQ articles about a joint cross-section of S&P 500 stock return data and related single-stock option data. Splitting the articles into a trading-time and an overnight archive, we distill tone from each of them. We show that media-expressed tone is informative about option markets and that both option data and tone predict stock returns. The predictive power of option variables is robust to partialling out tone, but varies depending on whether tone is from the overnight or the trading-time archive. A potential reason is that the archives differ in terms of their thematic content. Overall, we conclude that the informational content of option data for predicting single-stock returns extends beyond the information summarized in tone and traditional market factors.

Key words: media-expressed tone; option markets; stock return predictability; textual analysis; topic model;

JEL Classification: G12, G14, G41

[*] An earlier version circulated under the title "Textual sentiment, Option Characteristics, and Stock Return Predictability." We thank Elisabeth Bommes, Junjie Hu and Haowen Tang for excellent computational assistance.

[†] Adam Smith Business School, University of Glasgow, UK (*e-mail: CathyYi-Hsuan.Chen@glasgow.ac.uk*)

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

[§] Blockchain Research Centre, Humboldt-Universität zu Berlin, Germany; Sim Kee Boon Institute for Financial Economics, Singapore Management University, Singapore; Department of Probability and Mathematical Statistics, Charles University, Prague, Czech Republic; Department of Information Management and Finance, National Yang Ming Chiao Tung University, Hsinchu, China. (*e-mail: haerdle@hu-berlin.de*)

[¶] Lingnan College, Sun Yat-sen University, Guangzhou, China. (*e-mail: liuych26@mail.sysu.edu.cn*)

1 Introduction

It has been established – based on large bodies of text – that written documents contribute to price discovery in equity markets by carrying informational content that extends beyond the information sets created from past observations and other traditional market factors alone (for a survey, see Loughran and McDonald, 2016, and references therein). Text documents can contribute to price discovery, e.g., if processing textual information is costly in the sense of Hong et al. (2000) or due to behavioral biases of investors; see, e.g., Tetlock (2011). On the other hand, a growing number of studies stress the role of derivatives markets for price discovery in equity markets: Dennis and Mayhew (2002), Pan and Poteshman (2006), Xing et al. (2010), Stilger et al. (2016), among others, provide evidence that option market variables offer predictive power for future stock returns. The predictive power is attributed to the idea that informed traders maximize the value of their private information about stocks by trading in the derivatives market because option leverage and the relatively smaller number of market restrictions, such as short-sell constraints in stock markets, create powerful trading incentives.

In this work, we connect both strands of literature by examining the predictive power of single-stock option data for equity markets and measures of media-expressed tone simultaneously. Our motivation is that after all, when investors form their outlook for a particular stock based on textual information, they also need to choose a marketplace (the stock market or the derivatives market) to execute their trading idea. We therefore conjecture that textual information influences the equity market and the derivatives market alike. Moreover, a trading decision may rely on public information only, or on a mixture of private and public information. Hence, we separate textual information from other information embedded in option data and ask whether this "residual information" reaches beyond what is summarized in traditional market factors *and* textual information.

To accomplish this task, we measure firm-level textual news tone from a large text corpus scraped from NASDAQ news feed channels, which covers US companies listed in the S&P500 index. We refer to our textual measure as media-expressed "tone" rather than "sentiment," because we aim at setting it apart from the notion of sentiment as a not necessarily fact-based manifestation of emotions (Baker and Wurgler, 2007); see Section 2.2 for more details. We then explore whether trading-hour media tone is informative about three key single-stock option characteristics (OC-s), namely implied volatility (i.e., *IV*), out-of-the-money put prices (i.e., *Put*), and the implied volatility skew (i.e., *Skew*). Paralleling the findings from stock return data, we establish that both firm-level media-expressed tone as well as the cross-sectional aggregates of firm-level tone, i.e., tone indices, have a measurable impact on these OCs.

Equipped with this empirical evidence, we examine the predictive power of single-stock OCs for equity returns. In line with previous research, we find that OCs predict stock returns and remarkably that they continue to do so in the presence of tone variables, whereby the negative tone index emerges as a particularly powerful predictor variable. To study this predictive power more closely, we use the tone data along with conventional predictors to partial out publicly available information absorbed in option data. Using these orthogonalized components of OCs, we find that they still predict stock returns, which signals a substantial amount of insider information. Lastly, we check the economic significance of the statistical results and compare the profits of two long-short trading strategies. The first one – along with the extant literature – is based on OCs only, while the second one builds on the OCs orthogonalized to tone. We find that the latter strategy dominates the former in terms of Sharpe ratio, no matter which OC it is based on. In particular, the *Skew* residual-based strategy can obtain a Sharpe ratio of 3.93 (versus 2.87 for the *Skew*-based one), while it is 2.23 (versus 0.24) for *IV*, and 1.27 (versus 0.21) for *Put*. Thus, we conclude that the information content of option data reaches beyond what is summarized both in traditional risk factors and media-expressed tone.

In addition, we also discover new results about the divergent informational content of trading-hour versus overnight information. All our predictive stock return regressions underline that overnight information, i.e., information collected from articles in the preceding – not overlapping – night, is more informative than the younger trading-time tone. This parallels recent findings of Boudoukh et al. (2019), who conclude that overnight news is more informative about firm fundamentals than news that is released during the trading day. In order to shed further light on our observations, we apply topic models to the two archives, the trading-time and the overnight archive, to unearth their hidden thematic content. We find that trading-time and overnight articles cover noticeably different topics with little mutual overlap. These differing thematic emphases could contribute to the distinct predictive power of the different news archives.

As regards our techniques of textual analysis, we build on a more refined tool kit than traditionally used in the extant literature. Usually, rooted in a "bag-of-words" document model, one employs a dictionary-based counting process, a so-called lexicon projection; see, e.g., Cao et al. (2002), Das and Chen (2007), Schumaker et al. (2012), Chen et al. (2014), and Zhang et al. (2016). Bommes et al. (2020), however, observe that supervised learning algorithms trained on a phrase bank realize superior classification results because they achieve a surpassing explication of the linguistic sentence structure. Following these insights, we develop a supervised learning algorithm trained on the phrase bank of Malo et al. (2014) to predict sentence-level tone, but reserve all the tone variables which we derive from a traditional lexicon projection based on the Loughran and McDonald (2011) lexicon for robustness purposes.

The outline is as follows: Section 2 describes the text corpus and briefly presents the techniques used to quantify media-expressed tone. In Section 3, we study tone, option data, and return predictability. Section 4 provides robustness checks. Section 5 concludes, with an appendix offering all details on the data and tone measurement. The archive of NASDAQ articles is accessible from the authors upon request.

2 Data and sentiment distillation

2.1 Description of the text corpus

The NASDAQ news platform offers news and financial articles from selected contributors, including leading media such as Reuters, MT Newswires, RTT news and investment research firms such as Motley Fool, Zacks, and GuruFocus. Articles are classified into a number of categories

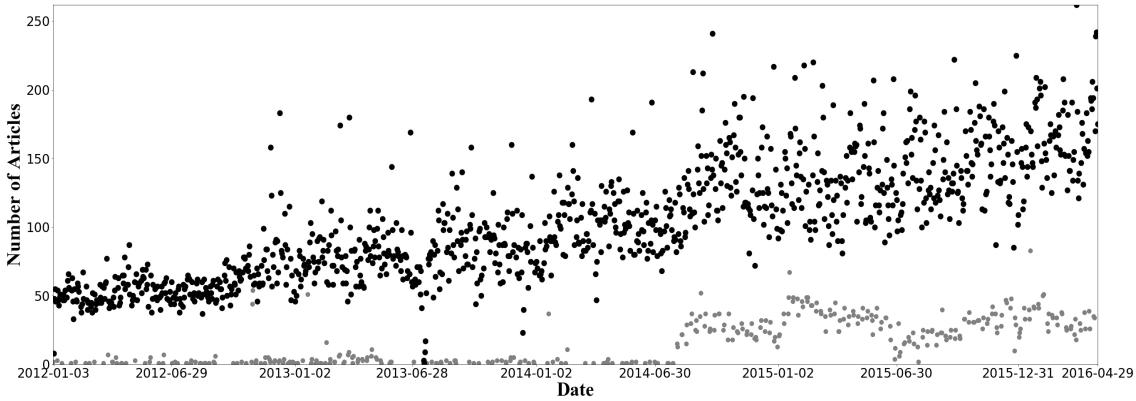

Figure 1: Number of article postings per day referring to the 97 companies listed in the S&P 500 index. A black point indicates the number of articles posted on a trading day, a gray point the number of articles posted on a non-trading day (weekend, holiday).

such as stocks, economy, world news, technology, fundamental analysis, etc. Articles in the stocks category are tagged with the ticker symbols of the stocks being discussed. Despite their origin, articles about companies not listed on NASDAQ are covered as well. We downloaded the time stamp, date, contributor, ticker symbols, title, and the complete text of 344,631 articles via an automatic web scraper written by Zhang et al. (2016). By intersecting the set of firms in our text corpus with the names in the IvyDB US database of OptionMetrics, we kept a subset of 119,680 articles featuring firms for which option data are available. In terms of firm universe, this is a pool of 97 firms, all of which are constituents of the S&P 500 and belong to nine industry sectors; for the full list, see Appendix A.1. As the sample period contains 1,581 calendar days, of which 1,088 are trading days, the 97 firms are receivers of approximately one piece of news per day.

In Figure 1, we illustrate 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 June 30, 2014, possibly due to the 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

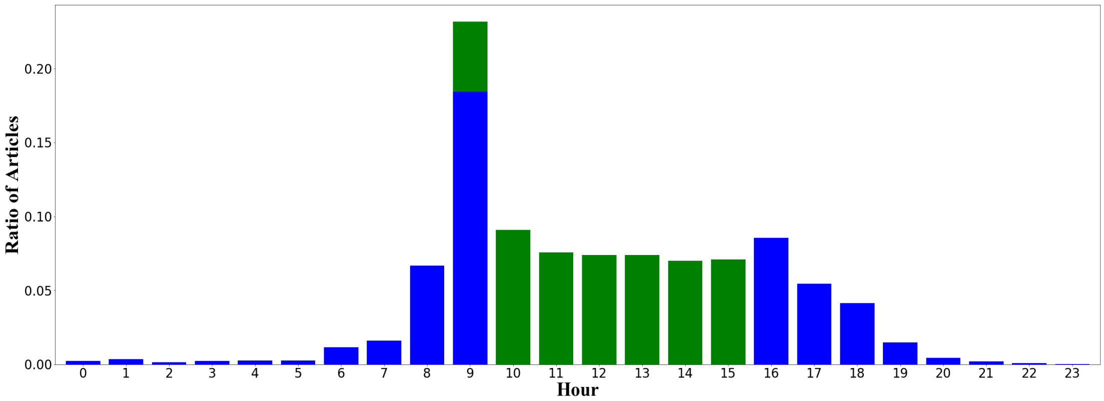

Figure 2: Hourly distribution (ET) of NASDAQ article postings. Hourly labels indicate the full hour, say from 08:00:00 a.m. to 08:59:59 a.m., etc. Blue indicates non-trading hours, green trading hours. Height of bar denotes the frequency of articles posted during that hour. The hour from 9:00:00 a.m.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

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. According to Figure 2, 33,160 articles (29.3%) are 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 the tradition of morning conferences within the finance industry. Moreover, 56,833 articles (50.3%) are found during active trading hours. The sample documents 23,087 articles (20.4%) 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.

2.2 Measurement of media-expressed tone

To quantify tone, we pursue two strategies, which we touch upon here (see the appendix for all details): (1) a lexicon approach; and (2), a more advanced supervised learning method. Both methods allow us to measure firm-level tone. We use the term “tone” to differentiate it from the more narrow notion of a “sentiment” as an irrational expectation (Baker and Wurgler, 2007,

p. 129). While our text corpus may contain statements in this sense of a sentiment, the topic analysis shows that the articles feature discussions about earnings, dividends, acquisitions, etc., which evidences clear traits of fundamental news. We therefore prefer the broader term "tone." More simply, we will also refer to our measures as "bullishness," although they just measure the degree of positive media-expressed tone.

In finance, lexicon projection techniques to measure textual tone were pioneered by Antweiler and Frank (2004), Das and Chen (2007), and Tetlock (2007) and they are still the most widely applied techniques to date. In lexicon projection, one compares the words of a document, where the term "document" can refer to a whole article or any substructure such as a sentence, with entries in a "lexicon." A lexicon, also called "dictionary," is a tabulated collection of words associated with certain attributes, such as a positive or negative polarity. Our dictionary of choice is the Loughran and McDonald (2011) lexicon, as it has been created specifically to parse financial news. Sentiment measurement then reduces to a tallying process, i.e., one counts positive and negative words in the respective document. Weighting and averaging yields a fraction of positive (negative) words per day per document, based on which one can assign a dominating polarity to the entire document and thus to a particular ticker symbol. If multiple firms are mentioned, we apply a slicing technique as in Wang et al. (2015); see Appendix B.2 for a precise account of all details of our implementation.

Lexicon projections have been criticized because of their mere "bag-of-words" concept, which strips the words from their contextual and semantic orientation. We therefore also developed a supervised learning method, which is able to incorporate the contextual semantic orientations. Instead of a lexicon, it requires a training data set, which we borrow from Malo et al. (2014). On this human-annotated financial phrase bank, we train a score-based linear discrete response model, which allows us to predict tone in the NASDAQ article database; see Appendix B.3 for all details on training and tone measurement.

Comparing the lexicon method (LM) with the supervised method (SM) on the training data set, we observe that the mean accuracy of the SM sentence-level method (with oversampling) is 80%, whereas the one based on LM only achieves an accuracy of 64%. A deeper analysis by means of the confusion matrix, which we report in Table 1, reveals that LM produces false negatives (type 2 error) and false positives (type 1 error) more often than the SM method does. Consequently, compared to lexicon projection, SM achieves a higher precision, i.e., finds more relevant results, than the irrelevant ones (one minus type I error). SM also has a higher recall, i.e., it retrieves more relevant results (one minus type II error). On these grounds, we proceed with our main analysis using the new SM method but still keep LM for reasons of robustness owing to its classical role in the literature.

2.3 Description of quantified measures of tone

After quantifying tone by the two methods, we construct – inspired from Antweiler and Frank (2004) – two firm-specific bullishness measures for each trading day: a trading-hour measure $B_{i,t}$ and an overnight measure $B_{i,t}^{on}$; see (9). 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; 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. Thus, trading tone on t is more recent than overnight tone on t. On a Monday, the overnight tone also covers the entire weekend starting on market close of the preceding Friday. This design aligns the date structure between the textual news channel and the option trading data. Summarizing, we work with the next variables of media-expressed tone:

(1) firm-specific bullishness $B_{i,t}$ $(B_{i,t}^{on}) \in [-1, 1]$ for the trading hour period (the overnight period): positive value of $B_{i,t}$ or $B_{i,t}^{on}$ implies positive tone 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 ($BN_{i,t}^{on}$ for the overnight period) where $\mathbf{I}(\mathcal{A})$ is the indicator function of the event $\mathcal{A}$;

(3) an aggregate index of tone $B_{idx,t}$ $(B_{idx,t}^{on})$ for the trading hour period (the overnight period) computed as an equally weighted cross-sectional average of $B_{i,t}$ $(B_{i,t}^{on})$; see the top panel of Figure 3;

(4) an aggregate negative tone index $BN_{idx,t}$ $(BN_{idx,t}^{on})$ for the trading hour period (the overnight

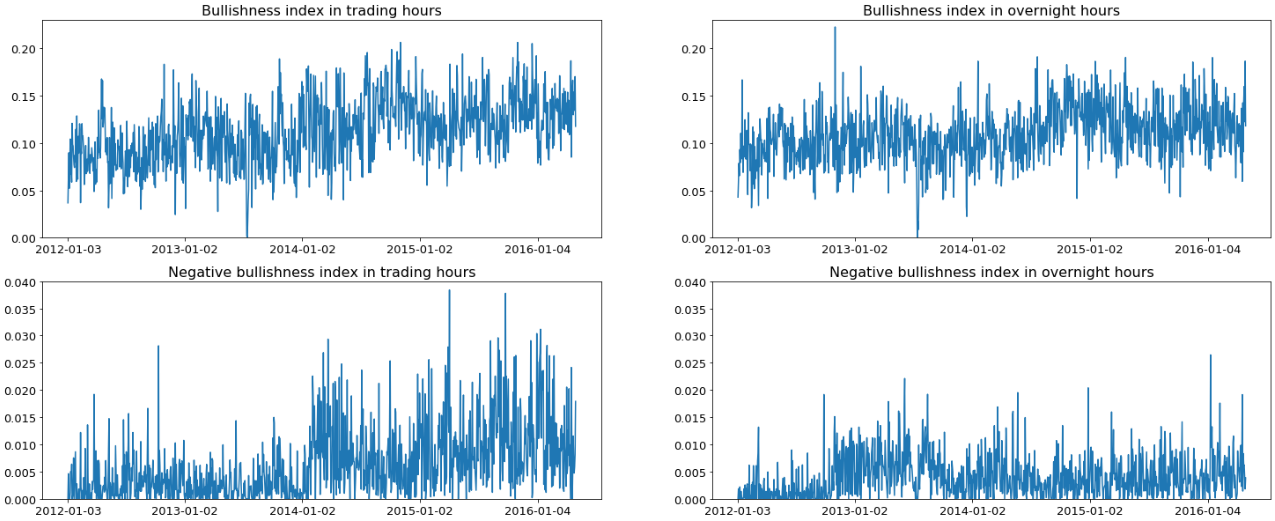

Figure 3: Daily bullishness index B_{idx} and B_{idx}^{on} (top panels) and the negative bullishness index BN_{idx} and BN_{idx}^{on} (lower panels), constructed during the trading hours' (left-hand panels) and the overnight hours' (right-hand panels) tone measures. The underlying tone measure is derived from the SM method. Source: NASDAQ articles, own computations.

period), computed as an equally weighted cross-sectional average of $BN_{i,t}$ ($BN_{i,t}^{on}$); see the lower panel of Figure 3.

Summary statistics of the data over all 97 firms are displayed in the upper panel of Table 2. Three important observations can be made. First, from the means and quantiles of the distributions of the means of BN_i and BN_i^{on}, it can be inferred that negative tone is more rare than positive tone. In part, this could be related to our sample ranging from Jan. 2012 to Apr. 2016. In the tone construction by means of supervised learning, we account for this fact by oversampling in the training process. Second, the statistical properties of tone gathered from the articles either during a trading day or overnight are similar. Our empirical analysis will investigate whether the two data sources are also similar in terms of economic content. Third, comparing SM-based tone measurements with those obtained from LM, we find much higher levels of the means of trading-time and overnight SM-bullishness, but a comparatively lower standard deviation (relative to the means) across the panels; in contrast, with negative bullishness, LM features higher levels of pessimism, albeit the standard deviation scales similarly. These differences are driven by two properties of lexicon projections. First, LM is domain-specific, i.e., LM requires a good match between the words of the text and the words collected in the lexicon. Hence the level of average tone one can find strongly depends on the quality of this match. Second, as Loughran and McDonald (2016) argue, the LM tends to measure positive tone less precisely than negative tone, because positive words tend to be ambiguous without any concomitant semantic orientation. This interpretation is further corroborated by the inferior classification results of LM; see Table 1 and the related discussion in Section 2.2. In Section 4, we study to which extent these statistical differences matter for the economics at work.

3 Tone, option markets, and stock return predictability

3.1 Measures of tone and single-stock option data

As a first step to analyze the relations between option markets, stock markets, and tone embedded in textual news flows, we ask whether news flow data are informative about option data. According to theory, there are two main mechanisms by which market tone can shape option prices: either via the pricing kernel, which summarizes the risk compensation a risk-averse investor requires for holding a risky asset, or via the physical transitional law of the asset price process; as an example, low media-expressed tone could tilt the pricing kernel, affect the conditional variance of the physical transition density, or both. Thus, we should be able to observe that option prices react to tone. Accordingly, we formulate

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

As set out in Section A.2, 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

$$OC_{i,t} = \alpha + \gamma_i + \beta_1 B_{i,t} + \beta_2^{\top} X_t + \varepsilon_{i,t} \,, \tag{1}$$

where $OC_{i,t} \in \{Skew_{i,t}, Put_{i,t}, IV_{i,t}\}$, $B_{i,t}$ is the quantified trading-time bullishness of firm i at time t – see (9) – and γ_i a firm fixed-effect; X_t is the vector of the Fama-French five factors included as control variables.

In regression (1), a potential endogeneity issue 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 on two grounds: either because an article is posted late on the NASDAQ servers or because it does not represent original information, but is written in response to a major piece of news published at an earlier time. As we could verify that the time lag between the release of an article and its publication on the servers is at most two hours, but often much less, we regard the first concern as minor; the second point, however, requires us to treat $B_{i,t}$ as an endogenous regressor.

We address endogeneity by means of a two-stage instrumental variable regression to estimate (1). As in the literature on dynamic panel models (Anderson and Hsiao, 1981), we choose the lagged tone $B_{i,t-1}$ as an instrument for $B_{i,t}$. This provides a valid exclusion restriction if the lagged instrument is determined before the error term at time t and the error term is independent of the past of all observed variables (conditionally on the covariates). This is a natural assumption in this context. Furthermore, according to basic option pricing theory, whether markets are complete or incomplete, options are priced at time t as the expected value of the payoff conditional on the time-t information set (Musiela and Rutkowski, 2006). This set necessarily includes contemporaneous media-expressed tone, if it matters at all for pricing options. A direct path from lagged tone to the option price, therefore appears not very plausible; in fact, all common option pricing formulae are functions of time-t variables only. Lastly, we include the Fama French five factors as covariates as they are widely acknowledged to capture a large mixture of market-wide information flows and therefore are established covariates in this context. In further but unreported robustness checks, we added further lags of the Fama French five factors and examined instrumentizing with $B_{i,t-2}$, all of which lead to the same conclusions.

In the reported regressions, the bullishness indices have been detrended to remove the slight trend apparent in Figure 3, and all bullishness measures are standardized to unit variance. This eases interpretations across the SM regression outputs. In Table 3, we report the first-stage results for H1, based on the trading-time SM-bullishness measure. Looking at column (1), we discover light but highly statistically significant patterns of autocorrelation of order one present in the bullishness measures, which we rely on for successful instrumentization. We test for underidentification by providing the Sanderson and Windmeijer (2016) χ^2 test and the Kleibergen and Paap (2006) rank-based Lagrange multiplier test. Both soundly reject underidentification. Because this is not sufficient for strong instrumentization, we additionally display F-statistics: the standard single equation F-test of excluded instruments, the Sanderson and Windmeijer (2016) F-test of excluded instruments, and Kleibergen and Paap (2006) Wald test. Appealing to the Staiger and Stock (1997) rule of thumb for interpretation, which considers the instruments weak when the F statistic is less than 10, we can conclude that instruments are indeed strong.

Proceeding to the second-stage results of H1 reported in Table 4, we see that H1 is strongly supported. More specifically, as positive news is released and bullish tone is formed subsequently, the skew of single-stock options becomes flatter; see column (1); moreover, OTM put prices decrease, column (4), and ATM implied volatility (IV) declines, column (7). Because a steep skew, high OTM put prices and high levels of IV are signs of agitated market regimes, we see that the positive tone expressed in NASDAQ articles has an appeasing influence on option markets.

Han (2008) studies whether three investor sentiment proxies, constructed from the Investors Intelligence's weekly surveys, the net position of large speculators in S&P500 futures, and the valuation errors of the S&P500 index influence S&P500 option prices. The results of Table 4 expand these findings in that firm-level media-expressed tone of a large text corpus is informative about single-stock option price data. This sheds further light on the price discovery role of single-stock option markets, which is often ascribed to leverage and built-in downside risk (Chakravarty et al., 2004). 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 tone on single-stock option price data. In Section 3.2, we distinguish furthermore between the informational content of OCs as reflected by tone, i.e., a public part, and a residual component, which captures private information.

Given the relation between firm-level OCs and firm-level tone, it is natural to ask whether individual OCs react to aggregate news. We do this in

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

H2 can be cast into the regressions

$$
\begin{aligned}
OC_{i,t} &= \alpha + \gamma_i + \beta_1 B_{i,t} + \beta_2 B_{idx,t} + \beta_4^\top X_t + \varepsilon_{i,t} \\
OC_{i,t} &= \alpha + \gamma_i + \beta_1 B_{i,t} + \beta_3 BN_{idx,t} + \beta_4^\top X_t + \varepsilon_{i,t}
\end{aligned} \tag{2}
$$

where $OC_{i,t} \in \{Skew_{i,t}, Put_{i,t}, IV_{i,t}\}$, γ_i is a firm fixed-effect, $B_{idx,t}$ is the trading-time tone index and $BN_{idx,t}$ the trading-time negative tone index as introduced in Section 2.3; the covariates X_t are as in H1.

We investigate H2 in a similar way by using in addition lagged tone indices as instruments. The first-stage results in columns (2)-(3) of Tables 3 again reject underidentification and weak instruments. The autoregressive structure in bullishness is once more corroborated, and – as one may expect – it is more pronounced among the indices. For interpretation, recall that $BN_{idx,t} > 0$ means negative tone. The second-stage results are supportive of H2. Column (2) of Table 4 shows that negative aggregate tone does not significantly increase the skew, while firm-specific bullishness still remains predictive of the skew. In contrast, according to column (3), the tone index B_{idx} flattens the skew, i.e., mitigates downside risk, but crowds out the firm-specific bullishness. Columns (5)-(6) of Table 4 show that the OTM put reacts negatively on firm-specific bullishness, increases with higher negative market-wide tone and decreases with higher market-wide tone, while the statistical significance of firm-specific bullishness is partially absorbed by the market-wide bullishness index. Similar patterns emerge for IV in columns (8)-(9).

Because all bullishness series have unit standard deviation, we can compare the size of the coefficients within a regression. Doing so, we see that coefficients of firm-specific bullishness are halved when the bullishness index is present; see columns (3), (6), and (9). This emphasizes the incremental informativeness of the market-wide bullishness index relative to firm-specific tone. Likewise, the informative content of the negative bullishness index is confirmed, with an exception in the skew regression in column (2).

Summarizing, we obtain solid evidence that media-expressed tone, both firm-specific and market-wide, is informative about price formation on derivatives markets. In unreported regressions, we also ran the very same specifications based on the overnight bullishness measures, instrumentized with lagged trading-time tone. We obtained very similar results, which underscores the discussion of this section. We therefore conclude that both trading-time and overnight textual tone is informative about single-stock option data.

3.2 Stock return predictability: option characteristics and tone

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 because traders, who possess a private information advantage over the public about firm fundamentals, execute their trading ideas in the option market. Subsequently, they profit from it as information diffuses in the market.

Given the evidence provided in Table 4, a natural question is, however, to what extent, if any, traders actually act on private information. Alternatively, trading ideas, which are inspired by the tone articulated in the NASDAQ articles, could only be executed via the option market. For this reason, we include both option characteristics and tone variables together in the predictive regressions of stock returns. If option characteristics are no longer significant when public media-expressed tone is controlled for, we can discount the importance of inside information implied in option characteristics. This motivates testing the following hypothesis:

H3: Tone contributes to stock return predictability on top of option variables.

We check H3 by

$$\begin{aligned} R_{i,t+1} &= \alpha + \beta_1 B_{i,t} + \beta_2 B_{idx,t} + \gamma OC_{i,t} + \theta^\top X_{i,t} + \varepsilon_{i,t} \\ R_{i,t+1} &= \alpha + \beta_1 B_{i,t} + \beta_2 BN_{idx,t} + \gamma OC_{i,t} + \theta^\top X_{i,t} + \varepsilon_{i,t} \end{aligned} \tag{3}$$

where $R_{i,t+1}$ denotes the return of firm i at time $t+1$ and $OC_{i,t} \in \{Skew_{i,t}, Put_{i,t}, IV_{i,t}\}$. We also replace the tone-related variables $B_{i,t}$, $B_{idx,t}$, and $BN_{idx,t}$ by their overnight mates $B^{on}_{i,t}$, $B^{on}_{idx,t}$ and $BN^{on}_{idx,t}$. The vector X_t comprises controls, which are typically added in predictive stock return regressions: the Fama-French five factors, the current return, idiosyncratic volatility, and market-wide volatility; see, e.g., Xing et al. (2010). We estimate a pooled regression because the F-test for fixed effects does not reject.

The results of (3) are reported in Table 5. We find that the skew predicts future returns with a negative sign; this indicates that the skew is a signal of future stock underperformance (Stilger et al., 2016); see columns (1) to (4). Columns (5) to (12) show that OTM put and IV are both significantly positive. Thus, the put price and IV 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 (Bollerslev et al., 2013, Chen et al., 2018). Beyond the evidence of Xing et al. (2010), this shows that the informational content of OCs is not annihilated by the media-expressed tone which one can extract from publicly accessible news servers such as NASDAQ's. These findings lend new support to the private information hypothesis on option data put forward in the literature and are further examined in Section 3.3.

As in the extant literature, all regression scenarios show that firm-specific trading-day tone $B_{i,t}$ is insignificant; see Tetlock (2007), Stambaugh et al. (2012), and Zhang et al. (2016). In contrast, the negative trading-day bullishness index has a clear directional impact on next day's returns: the higher the $BN_{idx,t}$, the lower the future return; see columns (1), (5), and (9). For the trading-day bullishness index $B_{idx,t}$, which includes both positive and negative tone, no predictive power is found; see columns (2), (6), and (10). Thus, the prediction power between average market-wide and market-wide negative trading-day tone is asymmetric and return prediction is only achievable in the presence of negative market tone. Theoretically, the predictability in states of low market tone can stem from various mechanisms. As argued in Diamond and Verrecchia (1987) and Engelberg et al. (2012), if short-sale constraints make trading more costly or even defer it, the speed of adjustment of security prices is reduced; this may lead to return predictability. Alternatively, Hong et al. (2000) show that analyst coverage is greater for stocks that are past losers; thus, negative information diffuses more gradually in the market, implying return predictability.

We turn to the predictive role of overnight tone. We find – as with trading-hour firm-level tone – no predictive power in firm-level overnight tone $B^{on}_{i,t}$; see columns (3), (4), (7), (8), (11), and (12). In contrast to trading-time information, the market-wide variables $B^{on}_{idx,t}$ and $BN^{on}_{idx,t}$, however, do both carry significant predictive power. Moreover, the coefficient on $BN^{on}_{idx,t}$ is about two times larger than that of $BN_{idx,t}$. To appreciate the economic magnitudes of the estimated coefficients, recall that tone variables have unit variance and that returns are measured in percentage terms; thus, if we adopt the average estimate across the columns for a back-of-the-envelope calculation, we find that a one-standard deviation move of negative trading-time bullishness impacts returns about -0.9 bp, whereas a one-standard deviation move in the overnight negative bullishness index is associated with a change of about -2.1 bp and in the overnight bullishness index with one of about $+3.0$ bp. In contrast, the effect of a one-standard deviation change in the skew amounts to about -1.7 bp ($\approx -0.50 \times 3.33\%$), in the OTM put to about $+4.9$ bp ($\approx 6.7 \times 0.73\%$), and in IV to about $+5.7$ bp ($\approx 0.55 \times 10.5\%$).[1] In summary, the overnight bullishness measures variables have an impact that is larger than that of skew and about half as large as the one of OTM puts and IV. The difference in coefficient size is marked between the tone indices of the trading-time versus the overnight archive.

To further ascertain this informational wedge, we investigate the topical content of the alternate archives in Section 3.4. Our evidence suggests that the archives have differing emphases in terms of topics covered. Independent of that, we find conclusive support of H3.

[1] For these approximations, we use the standard deviation of the respective OCs, which results when one ignores the panel-unit structure.

3.3 Stock return predictability: orthogonalized OCs and tone

In view of Section 3.2, we isolate the purported private informational component in OCs and provide further statistical and economic evidence of its existence. To this end, we carry out an anatomy of the "information content of option characteristics." By partialling out the public information and therefore operating on information orthogonal to tone-related information, we touch upon the fraction of unobserved information driving future returns. We study

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

H4 is concerned with the question of whether the tone condensed in $B_{i,t}$ can absorb the predictive power of OCs for future returns. This is checked by the panel regressions

$$\begin{aligned} R_{i,t+1} &= \alpha + \beta_1 B_{i,t} + \beta_2 B_{idx,t} + \gamma OC^{\perp}_{i,t} + \theta^{\top} X_{i,t} + \varepsilon_{i,t} \\ R_{i,t+1} &= \alpha + \beta_1 B_{i,t} + \beta_2 BN_{idx,t} + \gamma OC^{\perp}_{i,t} + \theta^{\top} X_{i,t} + \varepsilon_{i,t} \end{aligned} . \tag{4}$$

All variable definitions remain as stated below H3. The orthogonalized OCs, denoted by $OC^{\perp}_{i,t} \in \{Skew^{\perp}_{i,t}, Put^{\perp}_{i,t}, IV^{\perp}_{i,t}\}$, are the residuals of a regression of each $OC_{i,t}$ on firm-specific and market-wide tone and the controls. More specifically, we either regress on the trading-time measures or on the overnight measures, depending on which archive is studied. In consequence, $Skew^{\perp}_{i,t}$, $Put^{\perp}_{i,t}$ and $IV^{\perp}_{i,t}$ are orthogonal to public information embedded in the controls and to information ingrained in media-expressed tone of the respective archive.

Table 6 shows that we obtain similar results as reported in Table 5. As regards the tone variables, both the trading-time and the overnight measures, all results are confirmed. The informational wedge between overnight and trading-time measures of tone persists. Similar findings apply to orthogonalized OCs. For instance, $Skew^{\perp}_{i,t}$ enters into the equations with a negative coefficient as well and appears to be even a more precise measure of information: the p-values drop to about 1.4% as opposed to 0.6%. In all other dimensions, the results are almost identical to those reported previously and all OCs remain significant. In summary, public information-adjusted OCs still predict future returns; so does the market-wide tone, but the firm-level tone is not relevant. Taking the evidence regarding H3 and H4 together, we cannot refute the private information hypothesis.[2]

3.4 A topic analysis of the overnight and trading-time archive

The predictive stock return regressions of Section 3.3 suggest that there is an informational wedge between the articles of the trading-time and the overnight archive. A resolution to this observation could be a relationship between the news items' topics and their posting times. If topics or the thematic structures of information differ by archive, this may determine their relative informativeness. To investigate this hypothesis, we estimate a topic model on each archive. A topic model offers a probabilistic approach to discovering the hidden thematic structure in a text corpus by uncovering latent topics as distributions over sets of words. We opt for the Latent Dirichlet Allocation (LDA) of Blei et al. (2003), which is increasingly applied in empirical finance owing to its intuitive nature and excellent performance; see, e.g., Bao and Datta (2014), Brown et al. (2020), and Larsen et al. (2021).

In Table 7 (overnight archive) and Table 8 (trading-time archive), we report the top 15 most frequent words over 5 topics; see Section B.4 for model selection. In each exhibit, the top panel displays the topic words, while the lower panel documents the term probabilities, conditional on topic and topic assignment, which are indicative of the relevance of a word within a topic. Because the LDA is agnostic about the underlying economics, the topics need to be interpreted as the "principal directions" of the textual data clouds collected in the two archives.

Eyeballing the words identified as the most prominent ones, one observes that a number of words, such as *analyst, earning[s], revenue, investor,*[3] etc., coincide across the two archives. Because the

[2] An alternative interpretation of our findings would be that there is no private information embedded in option data, but that option traders are just better at reading and processing public information. While market segmentation is well known to hamper information flows and thus information processing, this interpretation appears scarcely plausible: The NASDAQ text corpus and its news content is produced by financial equity analysts and targets stock traders as audience; thus, it is not generated in an isolated market segment as is the case, for instance, in Addoum and Murfin (2019). We therefore dismiss this hypothesis.

[3] By using the square brackets [...], we denote parts of text dropped during text normalization.

genre of text is the same in both archives, we verify, by comparing the Hellinger distances between the topic distributions, that a good separation of the topic distributions is indeed achieved; see Section B.4 and Table 9. Moreover, invoking the concept of topic coherence, we can document a good semantic similarity of the topic words within each archive; see again Section B.4 for the details.

We now compare the topics reported in Tables 7 and 8 in more detail. Consider, for instance, topic 1 in the overnight archive, which has a frequency of 33.5%, and topic 5 in the trading-time archive, with a frequency of 20.5%. Both are the most and the second-most important topics in the respective folders, and they share *analyst* and *earning[s]* among the leading words. This could characterize both as "earnings" topics. The next most important topic words, however, give a different spin to each of them. For topic 1 of the overnight archive, words like *market, outlook, closes [at xxxx] point[s], update, acquisition, ipo, prospect* suggest that these are market commentaries on the current trading day, which may include a general market outlook as well as a review of ongoing acquisitions and pending IPOs, besides a discussion of earnings. It appears natural that terms like *acquisition, ipo* are cited prominently in the overnight archive because acquisitions and IPOs are likely to be repeatedly discussed over some time and have little in common with the daily trading business, except perhaps for the first time when discovering an acquisition or the first days of a new listing. In contrast, the trading-time earnings topic features additional words like *stock, earning[s], dividend, gain, profit.* It thus concentrates on specific names and their earnings (dividends, profits) and it has an apparent association to energy markets (*energy, oil*). Because many industries depend on energy markets, shocks to, e.g., the oil market tend to have an immediate implication for trading (Elyasiani et al., 2011); this may explain the appearance of these terms in the trading-time topics. Finally, note that topic 4 of the trading-time archive features *earning[s]* too, but within the bigram *earnings reaction.* Thus, in contrast to topic 5, topic 4 discusses the market reaction to reported earnings results, again a subject one would naturally expect in a trading-time archive.

As another difference between the two archives, trading-time topics often tend to contain keywords that are suggestive of topical updates or an activity, such as *rally* in topic 2, *update, alert,* *trading* as in topic 3, *option, earnings reaction, follow indicator* in topic 4, and *afternoon* in topic 5; remarkably, the term *(press) release* only appears in topics 1, 2, 5 of the trading-time archive. Furthermore, the term *tale [of the] tape*, with which comparative discussions of stocks are referred to, is also only found in the trading-day archive. On the other hand, the overnight corpus accentuates noticeably words that are evocative of a more general and fundamental type of analysis, such as *outlook, quarter, deal, business, plan, store, manager[s] want, acquisition, prospect, industry, sale.*

Thus, the two archives appear to be distinct not only in terms of the topic distributions, but also in terms of information type. The topic models suggest that the trading-day archive discusses immediate updates and short-term trading opportunities, whereas the overnight archive offers information of a more fundamental nature. This discrepancy could contribute to the informational wedge between trading-time tone and overnight tone. In fact, the observation that more complex information requires time to be absorbed and is strategically placed during market close has been documented in the accounting literature, e.g., in Berkman and Truong (2009) and Doyle and Magilke (2009). It also coincides with Boudoukh et al. (2019) who find that relevant news about firm fundamentals tends to be generated during overnight times rather than during trading-times. The yardstick in their study is return volatility. As regards information processing, Brunner and Ungeheuer (2020) report that overnight earnings surprises are associated with a high level of information acquisition activity on EDGAR servers, not only after the market opens, but throughout the entire trading day, particularly when associated with large abnormal returns. This observation is hard to explain if markets did absorb overnight information instantaneously. Our analysis, which documents an informational wedge between overnight and trading-time information, points in a like direction and thus extends the present findings to articles and analyses posted via the NASDAQ servers. We stress that our results are robust to using the lexical projections to distill tone – see Section 4.1.

3.5 Private information long-short trading strategy

To further investigate the economic significance of private information reflected in the OC-residuals $OC^{\perp}_{i,t}$, we design a long-short trading strategy. Indeed, if the $OC^{\perp}_{i,t}$ is an isolated component of private information, it seems reasonable to expect a trading strategy based on $OC^{\perp}_{i,t}$ alone to be superior than that 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 after the market closes on day t, from the fixed-effect panel data regression of the OC on the firm-specific and aggregate index of tone variables, as well as the control variables as in (3). Sort the 97 firms in descending order of these residuals and separate them into deciles. We use an in-sample period of three years before day t to calibrate the coefficients of the regression equations, which are then used to calculate the residuals of each firm on day t.

Step 2: Before market closing on day t, if OC is $Skew$ (IV or Put), we sell (buy) the group with the highest residuals and buy (sell) 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 for day $t + 1$, and rebalance the portfolio by repeating the above two steps for day $t + 1$. The three-year in-sample training period to determine regression coefficients is rolled forward.

We compare our strategy with the purely OC-based strategy. The latter is constructed by using the day t's OCs to sort the 97 firms and building up a long-short portfolio for the day $t+1$ similar to the one in the Step 2 above.[5] 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

[4] This is consistent with the predictive regressions as depicted in Table 6 where the coefficients of $Skew^{\perp}$ have negative signs, while those of $IV^{\perp}$ and $Put^{\perp}$ have positive signs.

[5] Such OC-based trading strategies are inspired by Xing et al. (2010) who developed a long-short portfolio trading strategy based on skew for weekly data.

0.07%. These figures are motivated by the findings reported in Edelen et al. (2013) on the bid-ask spread of liquid US stocks. On top of the results displayed, we also carry out robustness checks on different training samples and quintiles, which leave the results qualitatively unchanged.

Table 10 exhibits the annualized returns of the trading strategies for the case of zero transaction costs. The results are favorable. For all OCs, the residual-based strategy earns a much better Sharpe ratio. For the *Skew* residual-based strategy, we find an annualized Sharpe ratio of 3.93 (versus 2.87), for IV 2.23 (versus 0.24), for *Put* 1.27 (versus 0.21). Our results echo the remarkable predictability of skew itself as documented in Xing et al. (2010). In addition and more importantly, however, we can document the much better performance of a portfolio strategy based on the residuals of the skew stripped off public information and tone. Similar patterns are also observed for the other two OC variables. 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 tone. The Fama-French adjusted returns (alpha) underline furthermore that these results are not driven by common market factors.

When we consider transaction costs, the residual-based strategies still dominate: the Sharpe ratios are 2.52, 1.80, and 0.89 (*Skew*, IV, *Put*) for residual-based strategies versus 1.69, 0.10, and 0.00 for OC-based ones when the proportional transaction cost is 0.02%, while -0.32, 0.91, and -0.02 (*Skew*, IV, *Put*) versus -0.72, -0.19, and -0.44 for the transaction cost 0.07%. Tables are omitted for the sake of space but available upon request.

In summary, our results suggest that after public information and textual tone are filtered from OCs, their unexplained component remains informative about future stock returns in a trading case. Thus, we can attach to the isolated information in option data an economic value besides the purely statistical regression evidence. In practice, however, it may eventually be difficult to profit from this advantage when transaction costs are high.

4 Robustness

4.1 Lexicon projections

As we explained in Sections 2.2 and 2.3, the evidence of the confusion matrix in Table 1 suggests that our supervised learning method is the superior classifier. Lexicon projections are, however, the most widely applied technique to date. For this reason, it is insightful to inspect the validity of our analyses based on LM tone data. The regression outputs remain unreported but are available upon request.

The first-stage results are very similar. Underidentification is strongly rejected. The weak instrument test statistics are, with a few exceptions, slightly lower than those reported in Table 3 but still sufficiently large to suggest that instrumentization is good. As regards the second-stage results, we can confirm that LM tone is informative about OTM put prices and ATM IV; the signs are as in Table 3. Interestingly, the LM tone measures do not have any statistically significant influence on the skew. The skew is a difference between two IVs measured at different moneyness points. Consequently, an accurate signal is required to predict the reaction. It appears therefore plausible that the ambiguousness of LM tone is related to the poor classification strength of lexicon projections; see Table 1.

The predictive stock return regressions are fully consonant with those of Tables 5 and 6: firm-specific tone is irrelevant, and so is the trading-time bullishness index. In contrast, the negative trading-time bullishness index as well as both overnight indices do influence next-day's stock returns. Furthermore, as measured by coefficient size, there does not emerge a clear picture. The estimated coefficient of the LM negative trading-time bullishness suggests an impact of -1.8bp of a one-standard deviation move, which is twice as large as that of SM. In contrast, the LM negative overnight bullishness has an impact of only -1.3 bp (SM: -2.1 bp), while the LM overnight bullishness index is of similar magnitude with $+2.8$ bp (SM: $+3.0$ bp).

In summary, we can substantiate the results of Section 3. The economic content of the tone measures obtained by lexicon projection may be somewhat weaker, but overall it is similar to that of our supervised learning method. Most importantly, the informational wedge present between articles of the trading-time and the overnight archive is corroborated and thus independent of the actual method of measuring media-expressed tone.

4.2 Regressions based on attention

A large body of literature has observed that attention matters for investor decisions. Psychological research purports that this is because attention is a scarce cognitive resource (Kahneman, 1973). Due to their inability to process massive amounts of information, investors are attention-restricted, which may result in short-term mispricings (Hong and Stein, 1999, Hirshleifer et al., 2011). Using textual analysis, Zhang et al. (2016) find that news about high-attention firms diffuses at a different rate than that about low attention firms, signaling an asymmetric attention on news tone. We therefore study to what extent the predictive return regressions are affected if we distinguish between high and low-attention firms within the sample.

We measure attention for a firm by the fraction of references to a given ticker symbol divided by the total number of firm references measured over the entire archive. By taking the 25% and 75% quantile, we split the firms into a high and a low-attention group and repeat the predictive regressions from Sections 3.2 and 3.3. We focus the presentation on the regressions with OTM puts only because the conclusions are the same for skew and ATM IV.

We first discuss trading-time tone; see Tables 11 and 12. Compared with Tables 5 and 6, the negative trading-time bullishness index in the low-attention group (left panel) is no longer significant, as is the trading-time bullishness index. In contrast, in the high-attention group (right panel), the negative trading-time bullishness index remains a significant predictor. This is consistent with Zhang et al. (2016) and Hirshleifer et al. (2011) and suggests that the informational value of trading-time articles could partially be driven by firms that receive many references. Remarkably, the informativeness of tone measures in the overnight archive is not subject to the attention level: Regardless of attention, the overnight index measures of tone are relevant predictors. Moreover, the estimated coefficients of significant predictors are all of similar magnitude. This suggests that our results, overall, are not strongly driven by a particular subgroup in our sample.

In summary, two interpretations of the results are possible. Either the more fundamental and complex information is conveyed in the overnight archive, as is suggested by the topic analysis in Section 3.4; or investors are less attention-limited after the market closes, which helps them digest and process the information provided (Hirshleifer et al., 2011). Both interpretations underscore the informational wedge between the two archives.

5 Conclusion

We study the informational content conveyed about 97 S&P 500 firms from a huge assortment of NASDAQ articles. We split this text corpus into an archive of overnight articles and an archive of trading-time articles, explore their thematic composition by means of a topic model and distill their media-expressed tone. We find that media-expressed tone is informative about OCs and that OCs predict stock returns. Combining both results, we investigate the informational content of OCs for stock returns in the presence of media-expressed tone. We find that OCs still carry predictive content, which may be attributed to the private information embedded in option data. More importantly, when we isolate from option data the information explained by the text corpus the predictive power remains. Thus, both data sources, option data and tone data, do not fully overlap in terms of informational content. Whether tone matters, however, depends on whether tone is distilled from the overnight or the trading-time text archive. A topic model suggests that this could be due to the different thematic structures of the information provided. We add further support to our conclusions by showing that a trading strategy based on option information where tone and public information is partialled out yields higher Sharpe ratios than the standard strategy based on OCs alone.

Acknowledgement

Financial supports from National Natural Science Foundation of China under grants 71991474, 71721001, U1811462, and 72073148, the European Union's Framework Programme for Research and Innovation Horizon 2020 (2014-2020) under the Marie Skłodowska-Curie Grant Agreement No. 897277, the European Union's Horizon 2020 research and innovation program "FIN-TECH: A Financial supervision and Technology compliance training programme under the grant agreement No 825215 (Topic: ICT-35-2018, Type of action: CSA), the European Cooperation in Science & Technology COST Action grant CA19130 - Fintech and Artificial Intelligence in Finance - Towards a transparent financial industry, the Deutsche Forschungsgemeinschaft's IRTG 1792 grant, the Yushan Scholar Program of Taiwan, the Czech Science Foundation's grant no. 19-28231X / CAS: XDA 23020303, are highly appreciated.

References

Addoum, J. M. and Murfin, J. (2019). Equity price discovery with informed private debt, *Available at SSRN 2869452* .

Anderson, T. W. and Hsiao, C. (1981). Estimation of dynamic models with error components, *Journal of the American Statistical Association* **76**(375): 598–606.

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. (2007). Investor sentiment in the stock market, *Journal of Economic Perspectives* **21**(2): 129–152.

Bao, Y. and Datta, A. (2014). Simultaneously discovering and quantifying risk types from textual risk disclosures, *Management Science* **60**(6): 1371–1391.

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., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation, *Journal of machine Learning research* **3**(Jan): 993–1022.

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. (2020). Textual sentiment and sector specific reaction, *IRTG 1792 Discussion Paper 2018-043*.

Boudoukh, J., Feldman, R., Kogan, S. and Richardson, M. (2019). Information, trading, and volatility: Evidence from firm-specific news, *The Review of Financial Studies* **32**(3): 992–1033.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction, *Proceedings of GSCL* pp. 31–40.

Brown, N. C., Crowley, R. M. and Elliott, W. B. (2020). What are you saying? Using topic to detect financial misreporting, *Journal of Accounting Research* **58**(1): 237–291.

Brunner, F. and Ungeheuer, M. (2020). Information, trade, and salient returns, *Available at SSRN 2931547* .

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.

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* **93**: 21–32.

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

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.

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.

Elyasiani, E., Mansur, I. and Odusami, B. (2011). Oil price shocks and industry stock returns, *Energy Economics* **33**(5): 966–974.

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.

Hirshleifer, D., Lim, S. S. and Teoh, S. H. (2011). Limited investor attention and stock market misreactions to accounting information, *The Review of Asset Pricing Studies* **1**(1): 35–73.

Hong, H., Lim, T. and Stein, J. C. (2000). Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies, *The Journal of Finance* **55**(1): 265–295.

Hong, H. and Stein, J. C. (1999). A unified theory of underreaction, momentum trading, and overreaction in asset markets, *The Journal of Finance* **54**(6): 2143–2184.

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.

Kahneman, D. (1973). *Attention and effort*, Englewood Cliffs, N.J., Prentice Hall.

Kleibergen, F. and Paap, R. (2006). Generalized reduced rank tests using the singular value decomposition, *Journal of Econometrics* **133**(1): 97–126.

Larsen, V. H., Thorsrud, L. A. and Zhulanova, J. (2021). News-driven inflation expectations and information rigidities, *Journal of Monetary Economics* **117**: 507–520.

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.

Loughran, T. and McDonald, B. (2016). Textual analysis in accounting and finance: A survey, *Journal of Accounting Research* **54**(4): 1187–1230.

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.

Musiela, M. and Rutkowski, M. (2006). *Martingale Methods for Financial Modelling*, second edn, Springer-Verlag, Berlin, Heidelberg.

Newman, D., Lau, J. H., Grieser, K. and Baldwin, T. (2010). Automatic evaluation of topic coherence, *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, Association for Computational Linguistics, pp. 100–108.

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.

Pawara, P., Okafor, E., Groefsema, M., He, S., Schomaker, L. R. and Wiering, M. A. (2020). One-vs-one classification for deep neural networks, *Pattern Recognition* **108**: 107528.

Röder, M., Both, A. and Hinneburg, A. (2015). Exploring the space of topic coherence measures, *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, Association for Computing Machinery, New York, NY, USA, pp. 399–408.

Sanderson, E. and Windmeijer, F. (2016). A weak instrument F-test in linear IV models with multiple endogenous variables, *Journal of Econometrics* **190**(2): 212–221.

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.

Staiger, D. and Stock, J. H. (1997). Instrumental variables regression with weak instruments, *Econometrica* **65**(3): 557–586.

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. (2011). All the news that's fit to reprint: Do investors react to stale information?, *Review of Financial Studies* **24**(5): 1481–1512.

Wang, G., Wang, T., Wang, B., Sambasivan, D., Zhang, Z., Zheng, H. and Zhao, B. Y. (2015). Crowds on wall street: Extracting value from collaborative investing platforms, *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, Association for Computing Machinery, New York, NY, USA, pp. 17–30.

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.

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 97 companies

Apple Inc. (AAPL); AbbVie Inc. (ABBV)[6]; 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. (H-PQ); 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

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

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 Financial data

We match daily stock and option data to the measures of tone obtained from the text corpus. They include 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. 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.

We define moneyness 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 2.

In the regressions, we use daily Fama-French 5-factor data collected from K. R. French's website.[7]

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

B Methodological details of text analysis

B.1 Text normalization

Text normalization comprises preprocessing steps necessary before applying tools from textual analysis. The steps include: (1) tokenization, i.e., the process of breaking the text into word-based units called tokens; (2) converting uppercase letters to lowercase letters; (3) removing non-alphabetic characters and punctuation; (4) stop words removal, i.e., the removal of frequent words like *the, a, on*, etc., that do not carry any important meaning; (5) lemmatization, i.e., the process of reducing inflectional forms to a common base form based on a morphological analysis of words; (6) part-of-speech tagging, which assigns parts of speech to each word of the text (nouns, verbs, adjectives, etc.). Depending on the specific application, we then build n-grams, i.e., contiguous sequences of n items from the text, where $n = \{1, 2\}$. For this process, we rely on the Python module "Natural Language Toolkit" of Bird et al. (2009).

B.2 Lexicon projection

We illustrate lexicon projection for a positive tone Pos; the calculation is analogous for the negative tone Neg. Assume that the textual data only contain articles regarding one specific company i and 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 $N_{i,t} = \sum_{j=1}^{J} w_{i,t,j}$. Then one measures the positive tone using the fraction of positive words per day:

$$Pos_{i,t} = N_{i,t}^{-1} \sum_{j=1}^{J} \mathbf{I}\left(W_j \in L_{Pos}\right) w_{i,t,j} \;, \tag{5}$$

where L_{Pos} denotes the set of positive words in a predefined dictionary, here the Loughran and McDonald (2011) lexicon.

Eq. (5) is usually adjusted to account for negation: If the distance between a tone 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). Specifically, if L_{Neg} and L_{Pos} are the sets of negative and positive words, respectively, and additionally, $f_{i,t,j}$ and $u_{i,t,j}$ account, respectively, for the frequency of negated negative and negated positive words in $D_{i,t}$, we refine (5) to:

$$Pos_{i,t} = N_{i,t}^{-1} \sum_{j=1}^{J} \left\{ \mathbf{I}\left(W_j \in L_{Pos}\right)(w_{i,t,j} - u_{i,t,j}) \; + \; \mathbf{I}\left(W_j \in L_{Neg}\right) f_{i,t,j} \right\}. \tag{6}$$

Because a sentence level is often more precise (Wiebe and Riloff, 2005, Wilson et al., 2005), we work with a sentence-based polarity. Fix a company i and a date t, drop these indices for notational simplicity, and define (in abuse of the index j) as in (5) and (6) the positive/negative tone on the sentence level of a given document. We calculate for each sentence j, $j = 1, \dots, n$, its polarity as $Pol_j = \mathbf{I}(Pos_j > Neg_j) - \mathbf{I}(Pos_j < Neg_j)$ and finally aggregate them as

$$FP = n^{-1} \sum_{j=1}^{n} \mathbf{I}(Pol_j = 1) \tag{7}$$

$$FN = n^{-1} \sum_{j=1}^{n} \mathbf{I}(Pol_j = -1)\,, \tag{8}$$

where n is the number of sentences in the document. Eqs. (7) and (8) are the fraction of positive (FP) and negative (FN) polarity of company i at date t. Finally, we follow Antweiler and Frank (2004) and define

$$B_{i,t} \;=\; \frac{\log(1 + FP_{i,t}) \;-\; \log(1 + FN_{i,t})}{\log(2)} \tag{9}$$

as our measure of bullishness for company i on day t. Thus, $B_{i,t} < 0$ if the polarity of the text is negative, $B_{i,t} = 0$ indicates neutrality, and $B_{i,t} > 0$ suggests a positive polarity.

Because the articles we process are tagged with the underlying stock symbols, we can relate its textual content to a specific company. If in one article more than one company is referred to, we apply the slicing technique of Wang et al. (2015). If a firm i is mentioned in more than one document on date t, we set $B_{i,t}$ to the average over all computed measures. If a firm is not referred to at a given day, its tone is zero.

B.3 A supervised learning method (SM)

The basis of the machine learning approach is the financial phrase bank of Malo et al. (2014). Because the 5000 phrases were given to a group of 5 to 8 human annotators, who may disagree in their polarity judgment, we use only a sub-data set on which 66% of the annotators evaluating a particular sentence agree. Our Python code is described on http://www.quantlet.de in TXTfpbsupervised.

For numerization of the sentences, we employ 1-grams and 2-grams and create a word vector X from the vocabulary of the phrase bank, which has as entries the number of appearances of words in each sentence. We thus can define the score-based discrete response model $s(X) = \beta^\top X$, with parameter $\beta \in \mathbb{R}^p$ where p is large. Following Luhn (1957), the word matrix consisting of all sentences is then transformed into a *tf-idf* matrix. Since tone may be either negative, neutral or positive, we have a multi-class classification problem, which we solve with the one-versus-all approach (Pawara et al., 2020).

Given the training data $(X_1, Y_1), \dots, (X_n, Y_n)$ with $X_i \in \mathbb{R}^p$ and target $Y_i \in \{-1, 1\}$, and given the scoring function $s(X) = \beta^\top X$, we calibrate the predictive model via the regularized training error

$$n^{-1} \sum_{i=1}^{n} L\{Y_i s(X_i)\} + \lambda R(\beta) \,, \tag{10}$$

where $L(\cdot)$ denotes the loss function, $R(\cdot)$ a regularization term and $\lambda \geq 0$ a hyperparameter governing the amount of regularization. We use the Hinge loss $L(u) = \max(0,\, 1 - u)$ with the L_1-norm $R(\beta) = \sum_{i=1}^{p} |\beta_i|$. Optimization is achieved via the Stochastic Gradient Descent method. The regularization parameter was optimized using five-fold cross-validation. We oversampled sentences with positive and negative tone in the training set to obtain a balanced sample and control for the trade off between the type I and type II error.

From training, we obtain a vector $\widehat{\beta}$ with dimension $p \approx 43\,500$ from which we can predict tone in the NASDAQ article database. Each document is split up into its sentences and the corresponding score is calculated, yielding a predictor for the sentence polarity. We then aggregate sentence polarity into the fraction of sentences with positive and the fraction of sentences with negative polarity per article, company, and day. This allows us to compute our SM-based bullishness measure as in (9). We follow the same principles for multiple-firm references, multiple-article per firm citations as detailed in Section B.2. As a result, we obtain the bullishness measure $B_{i,t}$.

B.4 Topic model

The LDA of Blei et al. (2003) allows an article to feature multiple topics, while the overall number of topics is constant and fixed by the researcher. The latent topics are defined as distributions over word sets. The LDA decomposes the joint distribution of the observed words in the articles into a mixture of distributions of hidden random variables. The output is the conditional distribution of the hidden topic structure conditional on observed words; see Blei et al. (2003).

We apply the LDA to unigrams (nouns only) and bigrams separately to the overnight and the trading-time archive. We measure the degree of semantic similarity among the words using topic coherence; see Newman et al. (2010) and Röder et al. (2015). Words are "coherent", if they support each other in a certain context. We measure coherence by the average of normalized pointwise mutual information (NPMI, Bouma, 2009). NPMI is defined as $\text{NPMI} = -\log \frac{p(W_i,W_j)}{p(W_i)p(W_j)} / \log\{p(W_i,W_j)\}$, where W_i and W_j are topic words with marginal probabilities $p(W_i)$ and $p(W_i)$, respectively, and joint probability $p(W_i, W_j)$. Probabilities are estimated based on word co-occurrence counts. NPMI is bounded in $[-1, 1]$, where -1 indicates "never occurring together" and $+1$ "complete co-occurrence." Topic coherence is obtained by averaging NPMI over all topic words. Iterating from $K = 3, \dots, 20$ topics, we find the highest coherence for both corpora at $K = 5$, where the average NPMI is 0.65 and 0.62 for the overnight archive and the trading-time archive, respectively. These numbers indicate a good semantic similarity of topic words. Finally, we measure the distance between the topic word distributions based on the Hellinger distance, which is a metric and bounded in $[0, 1]$, which eases interpretation; see Table 9 for the definition. The distances are larger than 0.8; see Table 9, which suggests very good topic separation.

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

Pred / True	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	2193	2892	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 I error. Recall is a ratio of true positives to the sum of true positives and false negatives, equivalent to 1−type II error.

Table 2: **Descriptive Statistics**

Summary Statistics [%]						
	Variable	Mean	25%	50%	75%	Std
Supervised learning	$\overline{B}$	11.26	8.13	10.76	14.09	4.09
	$\overline{BN}$	0.63	0.30	0.43	0.89	0.44
	B_{idx}	11.26	8.82	11.26	13.57	3.39
	BN_{idx}	0.63	0.12	0.44	0.90	0.65
	$\overline{B}^{on}$	10.88	7.92	10.15	12.50	4.27
	$\overline{BN}^{on}$	0.39	0.18	0.31	0.53	0.32
	B^{on}_{idx}	10.88	9.06	10.80	12.62	2.87
	BN^{on}_{idx}	0.39	0.09	0.30	0.60	0.38
Lexicon projection	$\overline{B}$	1.12	0.00	1.42	2.74	2.71
	$\overline{BN}$	3.46	2.21	3.05	3.84	1.96
	B_{idx}	1.12	-0.57	1.08	2.77	2.44
	BN_{idx}	3.46	2.21	3.39	4.43	1.67
	$\overline{B}^{on}$	3.42	2.49	3.51	4.70	2.10
	$\overline{BN}^{on}$	1.83	1.04	1.56	2.18	1.22
	B^{on}_{idx}	3.42	1.65	3.34	5.10	2.54
	BN^{on}_{idx}	1.83	1.12	1.69	2.38	0.95
OC	$\overline{Skew}$	5.83	5.09	5.99	6.47	1.10
	$\overline{Put}$	0.57	0.30	0.42	0.68	0.47
	$\overline{IV}$	24.07	18.55	21.94	28.56	8.35

Descriptive statistics of tone for the supervised learning and the lexicon projection method and for the option characteristics (OC) over the sample period from Jan 2012 to Apr 2016, all expressed in %-terms. For data varying across panel units, we report the statistics on the respective panel means; e.g., $\overline{B}$ reports the means, standard deviations (Std), and quantiles on the 97 means of the daily bullishness measures; this calculation is applied to negative bullishness BN, both trading day and overnight, and the OCs, i.e., the implied volatility skew $Skew$, implied volatility IV and the relative put price Put. B_{idx} and BN_{idx} denote the respective bullishness indices over all 97 firms. Source: NASDAQ articles, IvyMetrics US (OptionMetrics), own computations.

Table 3: **First-stage and instruments' relevance**

	$B_{i,t}$	$B_{i,t}$	$BN_{idx,t}$	$B_{i,t}$	$B_{idx,t}$
	(1)	(2)		(3)	
$B_{i,t-1}$	0.0937	0.0941	0.0331	0.0873	0.0013
	0.000	0.000	0.000	0.000	0.660
$BN_{idx,t-1}$		0.0342	0.3410		
		0.000	0.000		
$B_{idx,t-1}$				0.0380	0.3107
				0.000	0.000
Controls	✓	✓	✓	✓	✓
SW χ^2 stat	259.8	256.0	163.4	219.0	262.0
p-val.	0.000	0.000	0.000	0.000	0.000
KP LM stat	66.0	65.7		62.4	
p-val.	0.000	0.000		0.000	
F stat	257.1	135.1	76.8	142.3	49.4
SW F stat	257.1	253.4	161.8	216.7	259.3
KP Wald stat	257.1	126.2		107.0	

The table reports the first-stage regressions for determining the influence of contemporaneous tone on OCs. We show the estimates and the p-value below. Tone is quantified by SM. Controls included are the Fama-French 5 factors. SW χ^2 is Sanderson and Windmeijer (2016) underidentification test, KP LM stat is the Kleibergen and Paap (2006) rank-based Lagrange multiplier test for underidentification. As regards weak instrument tests, F stat is the standard single equation F-test of excluded instruments, SW F stat is the Sanderson and Windmeijer (2016) F-test of excluded instruments, and KP Wald test is the Wald statistic owing to Kleibergen and Paap (2006). All statistics are computed from standard errors robust to heteroskedasticity and clustering. Number of observations is 105,183; number of ticker symbols is 97.

Table 4: **OCs and tone based on supervised method**

	$Skew_{i,t}$			$Put_{i,t}$			$IV_{i,t}$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$B_{i,t}$	-0.0027	-0.0032	-0.0010	-0.0013	-0.0019	-0.0006	-0.0187	-0.0260	-0.0083
	0.068	0.037	0.487	0.001	0.000	0.134	0.000	0.000	0.107
$BN_{idx,t}$		0.0016			0.0018			0.0239	
		0.225			0.000			0.000	
$B_{idx,t}$			-0.0029			-0.0012			-0.0184
			0.015			0.000			0.000
Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓
Fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓

The table reports the 2SLS results for determining the influence of contemporaneous tone on OCs by means of 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 $B_{i,t}$, $B_{idx,t}$, $BN_{idx,t}$. See Table 3 for first-stage results. Tone-related variables are quantified by SM. All regressions contain a constant and the Fama-French 5 factors as controls. In total, there are 105,183 daily observations, and 97 ticker symbols. All statistics are computed from standard errors robust to heteroskedasticity and clustering.

Table 5: **Predictive regressions with the OCs and tone**

	$R_{i,t+1}[\%]$											
						SM						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
$B_{i,t}$	-0.0050	-0.0046			-0.0033	-0.0030			-0.0034	-0.0031		
	0.437	0.488			0.607	0.645			0.603	0.637		
$BN_{idx,t}$	-0.0085				-0.0097				-0.0091			
	0.043				0.017				0.025			
$B_{idx,t}$		-0.0022				-0.0011				-0.0009		
		0.633				0.796				0.832		
$B^{on}_{i,t}$			-0.0024	-0.0061			-0.0004	-0.0041			-0.0001	-0.0040
			0.621	0.199			0.940	0.382			0.976	0.417
$BN^{on}_{idx,t}$			-0.0233				-0.0209				-0.0211	
			0.000				0.000				0.000	
$B^{on}_{idx,t}$				0.0299				0.0295				0.0301
				0.000				0.000				0.000
$Skew_{i,t}$	-0.4822	-0.4702	-0.5139	-0.5003								
	0.014	0.017	0.0090	0.0110								
$Put_{i,t}$					6.6762	6.6762	6.6762	6.8027				
					0.000	0.000	0.000	0.000				
$IV_{i,t}$									0.5560	0.5572	0.5486	0.5617
									0.000	0.000	0.000	0.000
Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
R^2 (%)	0.13%	0.13%	0.15%	0.16%	0.19%	0.18%	0.20%	0.21%	0.19%	0.19%	0.20%	0.22%

Tone-related variables are quantified by SM. All regressions include a global constant, Fama-French 5 factors, but no FE fixed effects (F-test indicates FE are jointly zero) and as controls the Fama-French 5 factors, current return, idiosyncratic volatility, and market-wide volatility. Below each estimate the p-value based on cluster-robust standard errors is displayed. Number of observations: 105,183; number of ticker symbols: 97.

Table 6: **Predictive regressions with the orthogonalized OCs and tone**

						$R_{i,t+1}$[%]						
						SM						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
$B_{i,t}$	-0.0048	-0.0044			-0.0048	-0.0044			-0.0048	-0.0044		
	0.457	0.504			0.452	0.500			0.455	0.503		
$BN_{idx,t}$	-0.0087				-0.0100				-0.0095			
	0.039				0.013				0.017			
$B_{idx,t}$		-0.0019				-0.0018				-0.0018		
		0.682				0.677				0.682		
$B^{on}_{i,t}$			-0.0024	-0.0061			-0.0023	-0.0061			-0.0023	-0.0061
			0.618	0.200			0.627	0.193			0.633	0.195
$BN^{on}_{idx,t}$			-0.0232				-0.0205				-0.0197	
			0.000				0.000				0.000	
$B^{on}_{idx,t}$				0.0295				0.0295				0.0295
				0.000				0.000				0.000
$Skew^{\perp}_{i,t}$	-0.5458	-0.5437	-0.5741	-0.5437								
	0.006	0.007	0.0040	0.0060								
$Put^{\perp}_{i,t}$					9.6379	9.5841	9.4257	9.5841				
					0.000	0.000	0.000	0.000				
$IV^{\perp}_{i,t}$									1.1176	1.1149	1.0984	1.1150
									0.000	0.000	0.000	0.00000
Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
R^2 (%)	0.13%	0.13%	0.15%	0.16%	0.22%	0.23%	0.24%	0.26%	0.26%	0.26%	0.27%	0.28%

Tone-related variables are quantified by SM. $Skew^{\perp}_{i,t}$, $Put^{\perp}_{i,t}$, and $IV^{\perp}_{i,t}$ are the residuals from regressing each variable on the tone variables of the respective archive and control variates. The reported regressions include a global constant, Fama-French 5 factors, but no FE fixed effects (F-test indicates FE are jointly zero) and as controls the Fama-French 5 factors, current return, idiosyncratic volatility, and market-wide volatility. Below each estimate the p-value based on cluster-robust standard errors is displayed. Number of observations: 105 183; number of ticker symbols: 97.

Table 7: **Topic Model Fit to Overnight Archive**

Topic	Prob.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.335	analyst	earning	market	investor	outlook	closes point	acquisition	update	manager want	ipo	thing	revenue	report	prospect	investment
2	0.153	share	service	deal	holding	quarter	plan	comment	value	partner	launch	talk	part	position	system	line
3	0.212	business	energy	portfolio	datum	territory	loss	result	rate	video	point	bank	demand	trade	firm	job
4	0.145	growth	dividend	store	sale	oil	surprise	etf	profit	fund	car	hike	move	asset	option	focus
5	0.155	stock	price	company	reason	buy	industry	drug	corporation	shareholder	move	closes point	issue	resource	decline	bond
1		0.285	0.079	0.024	0.021	0.011	0.011	0.010	0.009	0.009	0.009	0.008	0.007	0.006	0.005	0.005
2		0.076	0.029	0.027	0.025	0.023	0.020	0.017	0.017	0.014	0.014	0.010	0.010	0.010	0.010	0.009
3		0.038	0.035	0.034	0.033	0.030	0.021	0.020	0.020	0.015	0.015	0.015	0.012	0.010	0.010	0.010
4		0.073	0.067	0.033	0.032	0.026	0.024	0.021	0.017	0.016	0.012	0.012	0.010	0.010	0.009	0.009
5		0.200	0.029	0.024	0.016	0.015	0.012	0.010	0.010	0.010	0.010	0.009	0.007	0.007	0.007	0.007

Results of the Latent Dirichlet Allocation to the overnight archive. The best fitting model, selected by topic coherence, is presented. The top panel exhibits the topics based on their word distribution. The first column gives the topic number, the second column the topic frequency within the archive, and the next 15 columns the top 15 topic words/topical terms in decreasing order of occurrence. The lower panel tabulates for each topic word of the top panel its probability of appearance, conditional on topic and topic assignment.

Table 8: **Topic Model Fit to Trading-time Archive**

Topic	Prob.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.368	fund	revenue	commentary	system	business	result	etf	store	tale tape	deal	value	press release	investment	insurance	plan
2	0.127	growth	report	portfolio	guidance	commentary	tale tape	dollar	industry	record	press release	rally	semiconductor	contract	strategy	reason
3	0.128	update	share	trading	alert	line	holding	investor	options begin	tale tape	partner	thing	commentary	loss	territory	unit
4	0.172	option	history	earnings reaction	follow indicator	market	sale	buy	company	technology	outlook	bond	preferred stock	acquisition	issue	tale tape
5	0.205	analyst	stock	earning	energy	gain	dividend	afternoon	discount	oil	release	price	profit	datum	focus	resource
1		0.109	0.072	0.026	0.021	0.021	0.019	0.019	0.02	0.016	0.014	0.013	0.012	0.011	0.010	0.009
2		0.050	0.031	0.028	0.022	0.020	0.019	0.018	0.015	0.013	0.013	0.012	0.012	0.012	0.010	0.009
3		0.082	0.073	0.039	0.032	0.023	0.022	0.019	0.017	0.016	0.014	0.014	0.014	0.013	0.013	0.012
4		0.090	0.048	0.048	0.048	0.040	0.035	0.026	0.025	0.021	0.016	0.015	0.014	0.014	0.011	0.011
5		0.372	0.076	0.051	0.015	0.011	0.010	0.010	0.010	0.009	0.008	0.008	0.006	0.006	0.005	0.005

Results of the Latent Dirichlet Allocation to the trading-time archive. The best fitting model, selected by topic coherence, is presented. The top panel exhibits the topics based on their word distribution. The first column gives the topic number, the second column the topic frequency within the archive, and the next 15 columns the top 15 topic words/topical terms in decreasing order of occurrence. The lower panel tabulates for each topic word of the top panel its probability of appearance, conditional on topic and topic assignment.

Table 9: **Topical Distance Between Overnight and Trading Time Corpus**

		Trading time				
	Topics	1	2	3	4	5
Overnight	1	0.957	0.943	0.960	0.964	1.000
	2	0.907	0.873	0.902	0.919	0.954
	3	0.876	0.875	0.900	0.892	0.969
	4	0.920	0.891	0.922	0.878	0.986
	5	0.919	0.894	0.919	0.933	0.955

The distance between any two topic word distributions is measured by the Hellinger distance. Overnight topics are in rows, trading time topics are in columns. Given two probability distributions $P = (p_1, \dots, p_k)$ and $Q = (q_1, \dots, q_k)$, the Hellinger distance is defined as $H(P,Q) = 2^{-1/2}\sqrt{\sum_{i=1}^{k}(\sqrt{p_i} - \sqrt{q_i})^2}$.

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)	17.99	17.84	17.84	14.62	14.65	14.65
p-value	0.000	0.000	0.000	0.004	0.004	0.004
Ann. Return	0.55	0.56	0.56	0.43	0.44	0.44
Daily Vol. (in bp)	88.31			92.65		
Ann. Vol.	0.14			0.15		
Daily Sharpe Ratio	0.20			0.16		
Ann. Sharpe Ratio	3.93			2.87		
	IV residual			*IV*		
	Long-Short	FF_5	FF_3	Long-Short	FF_5	FF_3
Daily Return (in bp)	17.01	17.99	17.99	3.58	4.82	4.82
p-value	0.029	0.014	0.014	0.679	0.503	0.503
Ann. Return	0.49	0.57	0.57	0.06	0.13	0.13
Daily Vol. (in bp)	141.99			158.28		
Ann. Vol.	0.22			0.25		
Daily Sharpe Ratio	0.12			0.02		
Ann. Sharpe Ratio	2.23			0.24		
	Put residual			*Put*		
	Long-Short	FF_5	FF_3	Long-Short	FF_5	FF_3
Daily Return (in bp)	10.93	11.99	11.99	3.03	4.41	4.41
p-value	0.149	0.085	0.085	0.718	0.526	0.526
Ann. Return	0.28	0.35	0.35	0.05	0.12	0.12
Daily Vol. (in bp)	137.71			153.39		
Ann. Vol.	0.22			0.24		
Daily Sharpe Ratio	0.08			0.02		
Ann. Sharpe Ratio	1.27			0.21		

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 mean that the returns are adjusted by the Fama-French 5 factors and Fama-French 3 factors, respectively.

Table 11: **Attention-split regressions**

	$R_{i,t+1}$							
	Low attention				High attention			
	(1)	(2)	(3)	(4)	(4)	(5)	(6)	(7)
$B_{i,t}$	-0.0109	-0.0109			0.0008	0.0010		
	0.516	0.521			0.939	0.923		
$BN_{idx,t}$	0.0087				-0.0172			
	0.307				0.009			
$B_{idx,t}$		0.0008				0.0005		
		0.916				0.942		
$B^{on}_{i,t}$			-0.0037	-0.0084			-0.0103	-0.0123
			0.689	0.387			0.273	0.180
$BN^{on}_{idx,t}$			-0.0197				-0.0221	
			0.041				0.003	
$B^{on}_{idx,t}$				0.0365				0.0198
				0.000				0.017
$Put_{i,t}$	6.9849	7.0082	6.8862	6.982	12.0177	11.9357	11.6591	11.9637
	0.003	0.003	0.004	0.004	0.001	0.001	0.002	0.001
Controls	✓	✓	✓	✓	✓	✓	✓	✓
R^2 (%)	0.23%	0.23%	0.24%	0.27%	0.41%	0.39%	0.42%	0.41%
N	26086				26088			
Symbols	24				24			

Data set is split into low (25% quantile) and high-attention group (75% quantile). Tone-related variables are quantified by SM. The reported regressions include a global constant, Fama-French 5 factors, no FE fixed effects and the Fama-French 5 factors, current return, idiosyncratic volatility, and market-wide volatility as controls. Below each estimate the p-value based on cluster-robust standard errors is displayed. N is the number of observations, Symbols denotes the number of ticker symbol in each panel.

Table 12: **Attention-split regressions with orthogonalized OC**

	$R_{i,t+1}$							
	Low attention				High attention			
	(1)	(2)	(3)	(4)	(4)	(5)	(6)	(7)
$B_{i,t}$	-0.0128	-0.0127			0.0000	0.0004		
	0.446	0.454			0.999	0.963		
$BN_{idx,t}$	0.0085				-0.0162			
	0.317				0.013			
$B_{idx,t}$		0.0001				-0.0010		
		0.987				0.886		
$B_{i,t}^{on}$			-0.0061	-0.0109			-0.0117	-0.0136
			0.481	0.231			0.207	0.138
$BN_{idx,t}^{on}$			-0.0185				-0.0218	
			0.055				0.003	
$B_{idx,t}^{on}$				0.0370				0.0185
				0.000				0.026
$Put_{i,t}^{\perp}$	10.7570	10.7851	10.5644	10.7877	11.8651	11.8247	11.5488	11.9271
	0.011	0.011	0.013	0.012	0.019	0.019	0.021	0.019
Controls	✓	✓	✓	✓	✓	✓	✓	✓
R^2 (%)	0.27%	0.27%	0.28%	0.31%	0.35%	0.34%	0.37%	0.37%
N	26086				26088			
Symbols	24				24			

Data set is split into low (25% quantile) and high-attention group (75% quantile). Tone-related variables are quantified by SM. $Put_{i,t}^{\perp}$ is the residual from regressing tone and control variables. The reported regressions include a global constant, Fama-French 5 factors, no FE fixed effects and the Fama-French 5 factors, current return, idiosyncratic volatility, and market-wide volatility as controls. Below each estimate the p-value based on cluster-robust standard errors is displayed. N is the number of observations, Symbols denotes the number of ticker symbol in each panel.