Journal of Empirical Finance 18 (2011) 321–340

Contents lists available at ScienceDirect

Journal of Empirical Finance

journal homepage: www.elsevier.com/locate/jempfin

When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions☆

Axel Groß-Klußmann [a,⁎], Nikolaus Hautsch [b,1]

[a] Quantitative Products Laboratory, Humboldt-Universität zu Berlin, Germany
[b] Humboldt-Universität zu Berlin, CASE, CFS, Quantitative Products Laboratory, Germany

ARTICLE INFO

Article history:
Received 24 February 2010
Received in revised form 26 November 2010
Accepted 29 November 2010
Available online 7 December 2010

JEL classification:
G14
C32

Keywords:
Firm-specific news
News sentiment
High-frequency data
Volatility
Liquidity
Abnormal returns

ABSTRACT

We examine high-frequency market reactions to an intraday stock-specific news flow. Using unique pre-processed data from an automated news analytics tool based on linguistic pattern recognition we exploit information on the indicated relevance, novelty and direction of company-specific news. Employing a high-frequency VAR model based on 20 s data of a cross-section of stocks traded at the London Stock Exchange we find distinct responses in returns, volatility, trading volumes and bid-ask spreads due to news arrivals. We show that a classification of news according to indicated relevance is crucial to filter out noise and to identify significant effects. Moreover, sentiment indicators have predictability for future price trends though the profitability of news-implied trading is deteriorated by increased bid-ask spreads.

1. Introduction

Trading on financial markets is strongly influenced by public company-specific, macroeconomic or political information flows. Markets react sensitively to textual information updates—so-called "news"—which are announced on a regular and irregular basis. However, due to the enormous amount of news continuously released by modern electronic communication media nowadays it becomes increasingly difficult to process all news related to a certain financial asset. Particularly nonscheduled news is noisy and often hard to quantify and to interpret. It is not trivial to separate information from noise and to distinguish between relevant and less relevant news. Consequently, empirical studies typically focus on specific and easily identifiable news events such as scheduled macroeconomic announcements, political interventions or earnings announcements.

☆ For helpful comments and discussions we thank Rich Brown, Boris Drovetsky, Jakob Fiedler, Lada Kyj, Roel Oomen, two anonymous referees and the participants of workshops at Humboldt-Universität zu Berlin and at the Quantitative Products Laboratory. This research is supported by the Deutsche Bank AG via the Quantitative Products Laboratory and the Deutsche Forschungsgemeinschaft (DFG) via the Collaborative Research Center 649 "Economic Risk".
⁎ Corresponding author. Present address: Quantitative Products Laboratory (QPL), Alexander Str. 5, D-10178 Berlin. Tel.: +49 30 3407 5515.
E-mail addresses: axel.gross-klussmann@db.com (A. Groß-Klußmann), nikolaus.hautsch@wiwi.hu-berlin.de (N. Hautsch).
[1] Institute for Statistics and Econometrics and Center for Applied Statistics and Economics (CASE), Humboldt-Universität zu Berlin as well as Quantitative Products Laboratory (QPL), Berlin, and Center for Financial Studies (CFS), Frankfurt. Address: Spandauer Str. 1, D-10178 Berlin, Germany.

0927-5398/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jempfin.2010.11.009

This paper addresses the challenge of linking a virtually continuous and nonscheduled asset-specific news flow to intraday market activity. The fundamental objective of this study is to analyze to which extent high-frequency movements in returns, volatility and liquidity can be explained by the underlying mostly nonscheduled news arrivals during a day. To overcome the major difficulty of structuring and filtering news we employ the trading signals of an automated news engine. Such engines are technological innovations fueled by the algorithmic trading industry which computerize the interpretation of news based on linguistic pattern recognition techniques. The news engines are designed to provide signals on the meaning and the relevance of news items for future price movements as well as for future volatility and liquidity situations.

To our best knowledge, the present study is the first one systematically analyzing data from an automated news engine. We use the Reuters NewsScope Sentiment Engine which classifies firm-specific news according to positive, neutral and negative author sentiments based on linguistic pattern analysis of the respective news story. A further crucial feature of the engine is a numeric indicator classifying the relevance of news as well as a variable indicating the novelty thereof. Exploiting these numeric indicators of news sentiment, relevance and novelty we relate the firm-specific news to high-frequency returns, volatility, trading intensity, trade sizes, trade imbalances, spreads and market depth.

In specific, we aim to answer the following research questions:

(i) Are there significant and theory-consistent market reactions in high-frequency returns, volatility and liquidity to the intraday news flow?
(ii) Is trading on news-driven, machine-generated trading signals profitable?
(iii) Is the machine-indicated relevance of news empirically supported by corresponding market reactions?

Question (i) addresses a gap in the empirical finance literature which still lacks evidence on the impact of intradaily (nonscheduled) news on high-frequency market dynamics. Therefore, this study sheds some light on the question whether it is possible to empirically link linguistically pre-processed and filtered asset-specific news to the intraday trading process. We are particularly interested in the question whether there are market reactions beyond the effects induced by company-specific earnings releases which are well-known to have strong impacts. Hence, we explicitly discard all news on earnings announcement days and focus on effects which are predominantly driven by mostly nonscheduled and inhomogeneous news items. In this context, it is of interest to study not only reactions in returns but also in volatility, bid-ask spreads as well as liquidity demand and supply (represented by trading volume and market depth, respectively). Specifically the high-frequency news-driven effects on bid-ask spreads and market depth are widely unexplored. To our best knowledge, only Fleming and Remolona (1999) and Lee et al. (1993) report mainly summary statistics as some evidence of news-induced reactions in liquidity.

Research questions (ii) and (iii) are about the usefulness and effectiveness of machine-generated news feeds in intraday trading. While (ii) is addressed by testing for the significance of abnormal returns, question (iii) is investigated based on a classification of news into important and less important news items. The answers to (ii) and (iii) show to which extent linguistic analyses can help news vendors and traders to automatically structure the news flow. Finally, these questions provide also a first piece of evidence whether news engines have the potential to become building blocks in algorithmic trading strategies and thus major driving forces in market activity.

Compared to the vast literature on earnings announcements, only very few studies try to measure the market response to firm-specific intraday news. This is mainly because high-frequency news items are typically considered to be too noisy due to the interference with other sources of information. The work of Berry and Howe (1994) is an early attempt to link intraday market activity to aggregated measures of news like, e.g., the number of news. A similar approach is taken by Kalev et al. (2004) who document a positive relationship between the number of intraday news and stock market volatility. In an alternative intraday study, Busse and Green (2002) consider the impact of news released via television to test market efficiency. Ranaldo (2008) is the only analysis providing descriptive statistics on the impact of singular firm-specific news items on intraday trading processes. However, all studies show that the impact of news on intraday trading activity is only very weak and not identifiable anymore if earnings announcements are discarded. Moreover, typically news items have to be aggregated to reduce the influence of noisy and non-informative news. This is also confirmed by Mitchell and Mulherin (1994) reporting weak impacts of public news on a daily level. Finally, our study is also related to approaches based on the quantification of news texts. For instance, Tetlock (2007) and Tetlock et al. (2008) perform linguistic analyses of daily Wall Street Journal stories. Similarly, Antweiler and Frank (2004) link daily stock market activity to textual information from internet stock message boards. However, none of these approaches employ machine-processed and filtered textual news items.

Using 20 s aggregates of transaction data from 39 liquid stocks traded at the London Stock Exchange (LSE), we study news' impacts on abnormal returns, volatility, trading volume, average trade sizes, spreads, trade imbalances and market depth. While most studies analyze news effects based on fixed windows around the event dates, we model the complete underlying trading process. To avoid spurious regression results due to neglected dynamics and cross-dependencies between the variables, we employ a high-frequency Vector Autoregressive (VAR) model which is augmented by news-specific explanatory variables and explicitly accounts also for the naturally high proportion of zero variables arising from non-trading in a 20-second interval.

A major finding of our study is that high-frequency trading activity indeed significantly reacts to intraday company-specific news items which are identified as relevant. The fact that earnings announcements are discarded makes these results quite remarkable. We show that the observed market reactions well match theoretical predictions. By capturing dynamics and cross-dependencies in the VAR framework we find strongest effects for volatility and cumulative trading volumes. Bid-ask spreads, trade sizes and market depth do not necessarily directly react to news but indirectly through the cross-dependencies to volumes and volatility and corresponding spillover effects. Two findings confirm the usefulness of the linguistic pre-processing of news. Firstly, we find that the indicated sentiments have predictive contents for price movements around news arrivals. However, simultaneously rising spreads during these periods reduce the profitability of potential trading strategies. Secondly, only little market impact is found for news that is classified as not being relevant, while strong and significant effects are shown for relevant news. This result shows that news engines have the potential to successfully pre-structure news and to filter out noise.

The remainder of the paper is organized as follows. In the next section, we describe the underlying data set and present descriptive statistics. Section 3 reports empirical evidence for unconditional effects of published news items without explicitly controlling for time series dynamics and cross-dependencies in the processes. In Section 4, the econometric framework and corresponding results based on a high-frequency VAR model are given. Section 5 concludes.

2. Data

In order to facilitate the processing of new information, several news vendors offer software environments capturing particular characteristics of news in realtime. These tools electronically analyze available textual information using linguistic pattern recognition algorithms. Words, word patterns, the novelty of a news item, its type and other characteristics are translated into indicators of the relevance, novelty as well as of the tone of the item.

We use pre-processed news data from a news analytics tool of the Reuters company, the Reuters NewsScope Sentiment Engine. The data contain all 29,497 news headlines for the stocks we consider for 01/07 to 06/08 as observed on traders' screens. News arrival is recorded with GMT time stamps up to a millisecond precision. Each news message provides a sentiment, novelty and relevance indicator. The indicators are produced based on an automated linguistic analysis of news texts. The sentiment tags of the

Table 1
Descriptive statistics.

RIC	M. Cap in % of FTSE100	Money val. Traded	Return 1/7–6/8	Spread	Nr. of trades	Nr. of news	Relevant news	Rel. pos. news	Rel. neg. news	Rel. neut. news
AAL	2.41	58.44	0.37	2.41	85.87	1266	418	166	148	104
AV	1.36	84.21	−0.23	0.68	519.23	558	242	108	84	50
AZN	2.72	54.92	−0.19	1.43	175.34	861	391	142	170	79
BATS	1.90	38.33	0.31	1.40	232.87	312	131	81	29	21
BARC	3.08	391.72	−0.49	0.52	992.96	2243	874	390	331	153
BG	1.53	92.56	0.82	0.89	507.97	447	156	82	54	20
BGY	0.00	57.63	0.35	0.76	313.03	558	304	144	113	47
BLT	1.43	160.19	1.03	1.27	370.12	1300	279	122	100	57
BP	7.22	499.47	0.07	0.53	4546.40	2408	908	308	406	194
BSY	0.59	59.95	0.03	0.66	603.27	354	150	53	66	31
BT	1.61	251.73	−0.27	0.30	2097.74	508	170	85	51	34
CBRY	0.74	79.05	0.23	0.70	560.58	376	164	68	68	28
CNA	0.84	118.33	−0.18	0.34	855.75	348	134	66	48	20
DGE	1.76	64.39	−0.02	0.96	530.28	282	110	67	30	13
EMG	0.64	90.51	0.18	0.73	488.34	174	85	40	23	22
FP	0.30	134.54	−0.44	0.25	1134.61	469	223	110	77	36
GSK	5.00	131.54	−0.17	1.13	855.07	1219	408	127	148	133
HBOS	2.75	170.43	−0.65	0.81	650.74	927	320	99	155	66
HSBA	6.96	349.53	−0.09	0.56	1806.28	1986	695	293	224	178
III	0.30	22.21	−0.12	1.26	177.50	272	62	33	14	15
IMT	0.88	20.49	−0	1.98	97.36	397	207	118	60	29
ITV	0.27	175.74	−0.45	0.17	3149.78	378	190	53	118	19
LLOY	2.08	220.87	−0.33	0.49	1175.85	528	181	71	74	36
LSE	0.00	10.84	−0.22	2.67	99.64	700	281	144	73	64
MKS	0.78	105.32	−0.47	0.60	565.14	475	152	49	79	24
NG	1.29	65.86	0	0.67	499.17	312	124	56	38	30
NXT	0.26	22.78	−0.36	1.90	97.74	323	153	65	77	11
PRU	1.10	111.43	−0.05	0.68	512.25	417	173	88	40	45
RBS	4.05	450.73	−0.89	0.69	1046.44	1859	560	234	216	110
RIO	1.80	59.42	1.20	3.21	70.75	1223	362	199	87	76
RR	0.52	81.26	−0.06	0.54	502.99	397	238	125	39	74
SAB	1.14	36.72	0.10	1.41	254.56	368	181	98	45	38
SBRY	0.45	90.66	−0.14	0.54	718.96	652	307	169	92	46
SL	0.40	43.67	−0.16	0.45	454.73	321	189	82	77	30
STAN	1.33	54.13	0.24	1.45	213.59	735	328	173	96	59
TSCO	2.07	213.45	0.02	0.33	1034.07	506	190	95	55	40
ULVR	1.20	41.52	0.17	1.34	299.38	330	108	61	29	18
VOD	4.82	1426.32	0.15	0.14	508.25	1700	559	315	137	107
XTA	1.55	56.23	0.56	2.80	76.53	1008	326	164	120	42
Sum	69.16					29497	11033	4943	3891	2199

Note: RIC denotes the Reuters Identifier Code. The second column is the % market cap., defined as the shared price times the number ordinary shares in issue for 2007, as fraction of the market cap for the FTSE100. Money value (traded) is computed as the trade size times the respective price (traded turnover total in 01/2007 to 06/2008 in 100,000). Return refers to the % price change from 01/03/07 to 06/01/08. Spread and Nr. of trades (in 100,000) are averages per trading day. The News column refers to the number of news items per firm without overnight news and duplicate entries. Relevant news items are classified to be the ones with a relevance indicator equal to one. Rel. pos. news and Rel. neg. news give the numbers of relevant positive and negative items, respectively.

news are coded +1, 0 and −1 for a positive, neutral and negative tone of the underlying story, respectively. Relevance is given by a number in the [0,1] interval. The novelty indicator reflects how many news with similar content have been published prior to a certain news item. It can thus be used to identify initial news topics (novelty = 0) and updates on a topic (novelty > 0).

We select 40 stocks traded at the LSE which are most active in terms of the number of published news items. As we require data availability for 355 trading days from 01/03/2007 to 06/01/2008, the sample is ultimately cut down to 39 stocks. The fact that the selected stocks are also very actively traded allows us to study market dynamics based on a high frequency. Covering 70% of the market capitalization of the FTSE100, our sample can be considered as being representative for the FTSE100 (see Table 1).

The underlying transaction data is aggregated to 20 s intervals. We consider this aggregation level to be a good compromise between exploiting a maximum of information on the one hand and making the analysis still computationally tractable (given 1.5 years of data). To reduce the impact of market opening and closing effects, we discard the first ten and last ten minutes of a trading day. Intraday returns, volatility and liquidity are captured by the following variables computed over 20 s intervals:

(i) money value traded, defined as trade sizes in the intervals weighted by the corresponding mid-quotes,
(ii) average trade size, defined as the cumulated trade size divided by the corresponding number of trades per interval,
(iii) bid-ask spread, evaluated at the endpoint of each interval,
(iv) mid-quote returns over each interval,
(v) depth, defined as the volume pending at the best bid and ask level, evaluated at the endpoint of each interval,
(vi) volatility, defined as the sum of squared mid-quote transaction returns over each interval,
(vii) trade imbalance, defined as the difference of cumulated sizes of buyer and seller-initiated trades (identified by the Lee and Ready (1991) algorithm), normalized by the cumulated trade size,
(viii) absolute trade imbalance.

As shown by Fig. 13 (see Appendix A), all volatility and liquidity variables exhibit pronounced intraday trading patterns. To capture the seasonality, we standardize all processes by the yearly average of the corresponding underlying 20 s interval. We compute standardized variables according to

$$x^*_{jd} := \frac{x_{jd}}{1/n \sum_{d=1}^{n} x_{jd}},$$

where j denotes the specific interval of the trading day d and x represents the corresponding variable.

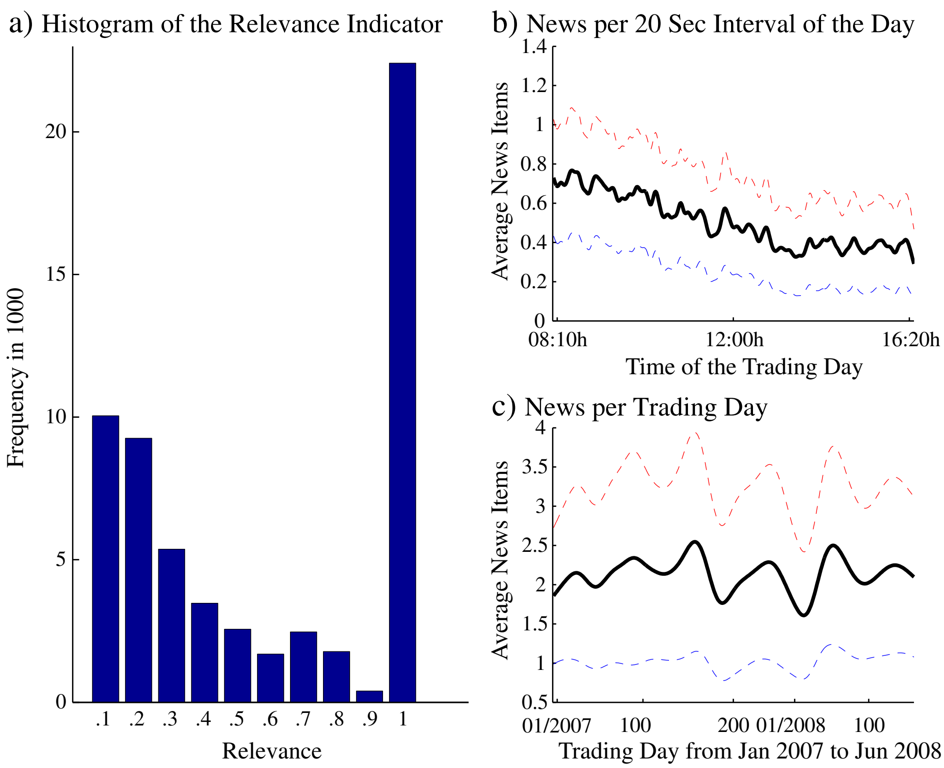

Fig. 1. a) Distribution of the relevance indicator. b) Distribution of news over a day and c) over the time span January 07 to June 08 (averages). Confidence bounds dotted. Smoothed via kernel regression.

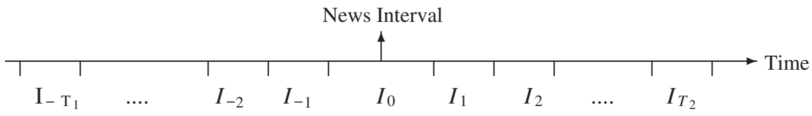

Fig. 2. Intervals around news arrival.

We delete all news on days of earnings announcements to minimize the influence of scheduled earnings releases on the results. This allows us to focus on the yet unexplored data of widely unscheduled intraday news driven mostly by random events. In addition, we only consider the news flow within a trading day and do not examine overnight news. Incorporating the latter would considerably increase the complexity of the study.

After pre-filtering, the number of news range from a minimum of 174 to a maximum of 2408 disclosures per stock for the 01/2007 to 06/2008 period (see Table 1). We observe that news tend to cluster in the first half of a day. As shown by Fig. 1 b), the news intensity peaks at the beginning of the trading period and decreases during the rest of the day. Fig. 1c) reveals that there is no pronounced yearly pattern of news arrival.

In order to identify potentially market-moving news, we distinguish between relevant and less relevant news according to the linguistic pre-analysis. Since we expect the reported relevance tag of news to be a relatively noisy measure, we classify items with an indicator value at (below) the maximum 1 as relevant (irrelevant) news (see Fig. 1a)).

3. Unconditional effects of news items

3.1. Impact on volatility and liquidity

Quantifying the unconditional impact of news without controlling for market dynamics and cross-dependencies between variables already provides important insights and serves as a basis for the econometric modelling in Section 4. We analyze 720 20-second intervals around the arrival of news items capturing 180 intervals before each disclosure and 540 thereafter.

Fig. 2 illustrates the timing of the intervals. I_0 denotes the specific 20-second interval around the news item, whereas T_1 and T_2 are the numbers of intervals before and after the disclosure, respectively. For each stock, we compute the average market reaction and corresponding standard errors over all event windows. For sake of brevity, we refrain from showing results for individual stocks but report pooled averages over the cross-section of stocks. Correspondingly, by denoting the market reaction of variable X to news item i during interval I_j as X_{iI_j}, the pooled average across all news events and all stocks is computed as $\overline{X_{I_j}} = 1/n\sum_{i=1}^{n} X_{iI_j}$, where n is the total number of news for all stocks. Given that the stocks have quite similar empirical characteristics (see Table 1), this proceeding allows us to highlight the results common to all stocks. Assuming (approximative) normality, the 95% confidence intervals of $\overline{X_{I_j}}$ are computed as two times the standard errors of $\overline{X_{I_j}}$. Since these standard errors reflect variations across all event windows as well as across the market, they capture overall news responses and statistical confidence thereof. Two robustness checks underscore the validity of the inference. First, the confidence intervals closely match those obtained from a parametric bootstrap. Second, to account for the fact that stocks with a high number of news naturally have a stronger weight in $\overline{X_{I_j}}$, we perform a robustness check using a group-means estimator instead of a pooled average. The corresponding results are qualitatively identical.[2]

Figs. 3–5 show the money value traded, realized volatility, bid-ask spreads, market depth, average trade sizes and absolute trade imbalances around relevant and less relevant news items. Note that by construction of the seasonality adjustment the mean of each series equals one.

The following findings can be summarized: First, we identify significant upward movements in money value traded, average trade sizes and volatility around the releases of news items. Hence, volatility and trading activity clearly increase when news are published. This supports a 'micro-foundation' of the mixture-of-distribution hypothesis as postulated by Clark (1973) and put forward by Tauchen and Pitts (1983) and Karpoff (1986). In this framework, price changes are essentially driven by trading on pieces of news, whereas uninformed traders tend to trade when they see large price movements. In consequence, the theory indeed predicts co-movements of volatility and volume. The observed effects are also well supported by market microstructure theory providing several explanations for higher trading activity during news announcements: (i) Larger trade sizes due to execution by better informed market participants according to Easley and O'Hara (1987), (ii) increased trading due to news-induced information asymmetry among market participants as advocated in the models of Kim and Verrecchia (1991) and Kim and Verrecchia (1994), (iii) trading because of differences in opinion of traders on news' topics as in Harris and Raviv (1993) as well as Kandel and Pearson (1995), and (iv) trading as a consequence of the attention grabbing behavior of investors as documented by Barber and Odean (2008). Beyond overall increases in volumes, we also observe slight increases in absolute trade imbalance reflecting that trading activity on the two sides of the market tends to become also more asymmetric in periods of information dissemination.

[2] See Appendix A for more details on the computation of standard errors.

Second, the release of a news item significantly increases bid-ask spreads but does not necessarily affect market depth. Hence, liquidity suppliers predominantly react to news by revising quotes but not by offered order volumes. This is well supported by asymmetric information based market microstructure theory (see, e.g., Easley and O'Hara (1992)) where specialists try to overcompensate for possible information asymmetries. Though on an electronic market there are no designated market makers, the underlying mechanism is similar: Liquidity suppliers reduce their order aggressiveness in order to avoid being picked off (i.e., being adversely selected) by traders which are better informed. For earnings announcements, such effects are also reported by Krinsky and Lee (1996).

Third, the machine-indicated relevance of a news item is clearly supported by corresponding market reactions. All variables respond significantly stronger if information is indicated to be of highest relevance. Actually, for less relevant news we cannot identify significant deviations of the analyzed trading variables from their pre-news levels. This finding is economically in line with Blume et al. (1994) who argue that higher volumes reflect a higher quality of news signals. Moreover, it shows that an appropriate filtering and structuring of the news flow (as provided by the news engine) is crucial to identify systematic market responses. In fact, the noisiness of less relevant news items is the major reason for the yet missing empirical evidence on statistically significant relationships between intraday news flow and high-frequency market activity.

Fourth, for most variables, above-average activities start already more than sixty minutes before the item arrival. This phenomenon is also known in case of periodically scheduled earnings releases. According to the model in Kim and Verrecchia (1994), trading prior to news depends on the degree of information leakages. Our results show that some market participants seem to have additional and partly more timely channels of information. Besides information leakage we attribute the pre-news reaction to a clustering of news. The clustering is an effect inherent in the production of news stories: Beginning with an alert about the news content, subsequent updates ultimately culminate in a full-blown story. Typically, the time between updating steps is small. The novelty indicator of the news data allows us to separate between 'news' (in its true sense) and updates on the topic published later. Fig. 6 shows money value traded and bid-ask spreads around 'initial' news and subsequent updates. Most strikingly, we find that trading on updated news is much more pronounced than trading on the initial news which strongly supports the notion of news clustering. Hence, market reactions become stronger if signals on news are repeated, refined and possibly enforced. This confirms the importance of accounting for the full history of news. Given that later published full stories are more precise than initial alerts, the findings support also the theoretical model of Tetlock (2010) who argues that the magnitude of the volatility and volume change at public news disclosures are positively related to the precision of the news.

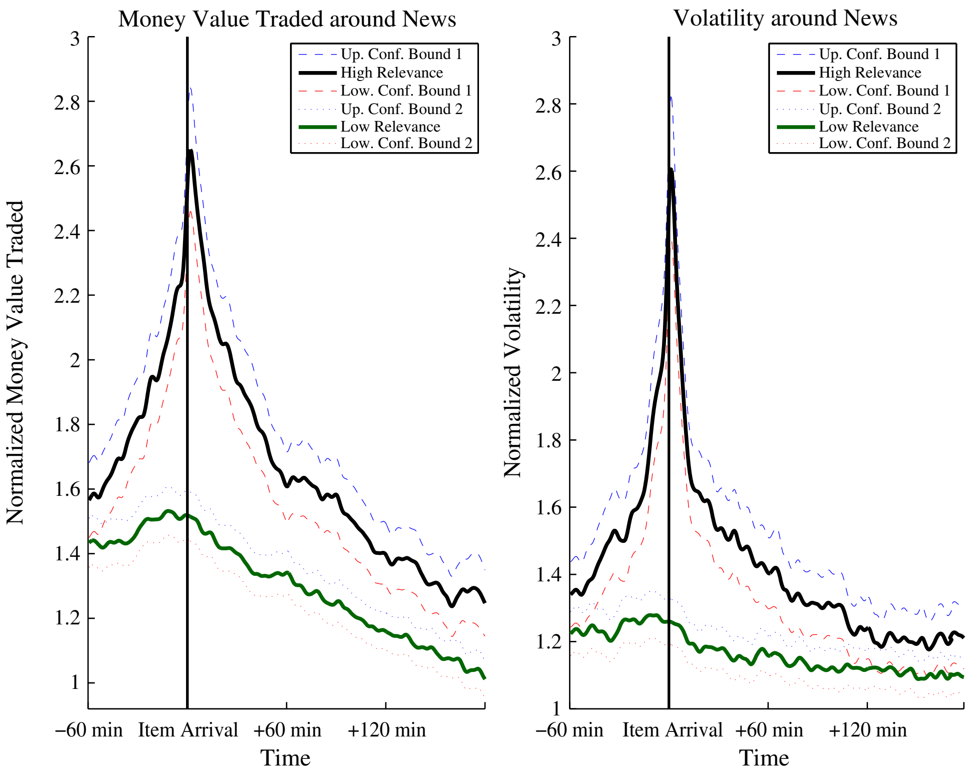

Fig. 3. Money value and volatility around news arrivals. Smoothed via kernel regression.

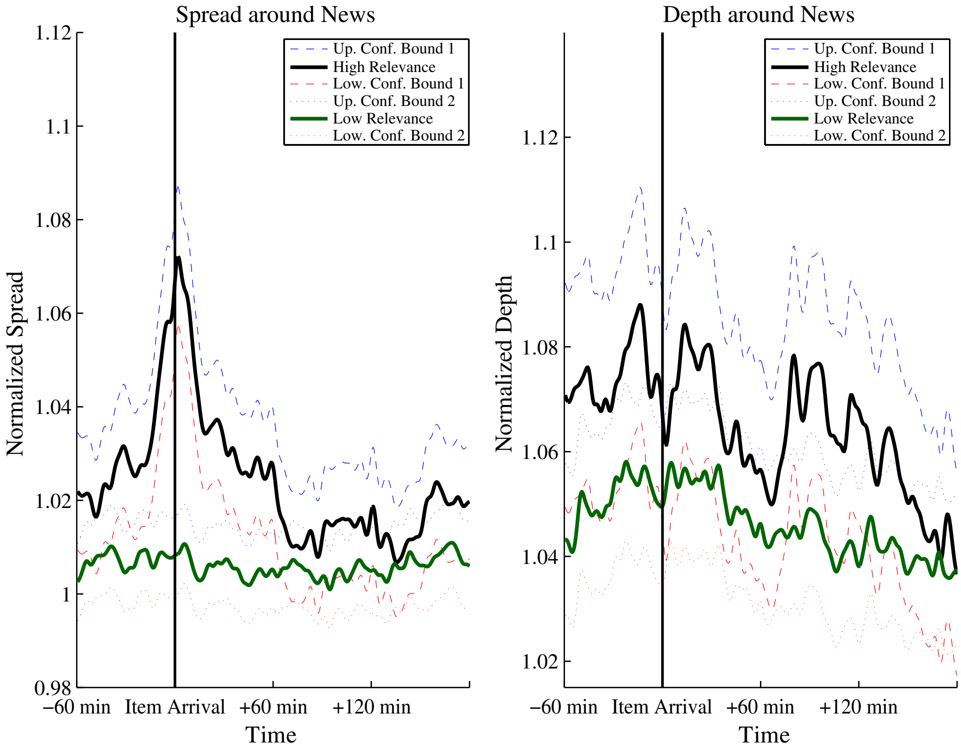

Fig. 4. Spread and depth around news arrivals. Smoothed via kernel regression.

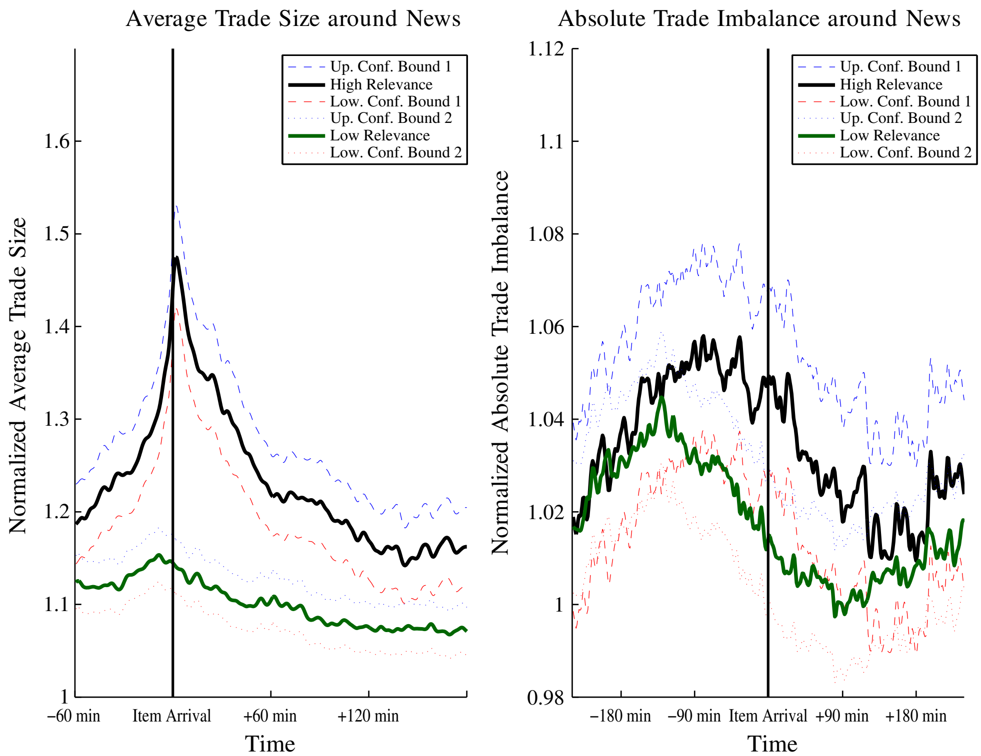

Fig. 5. Average trade sizes and absolute trade imbalance around news. Smoothed via kernel regression.

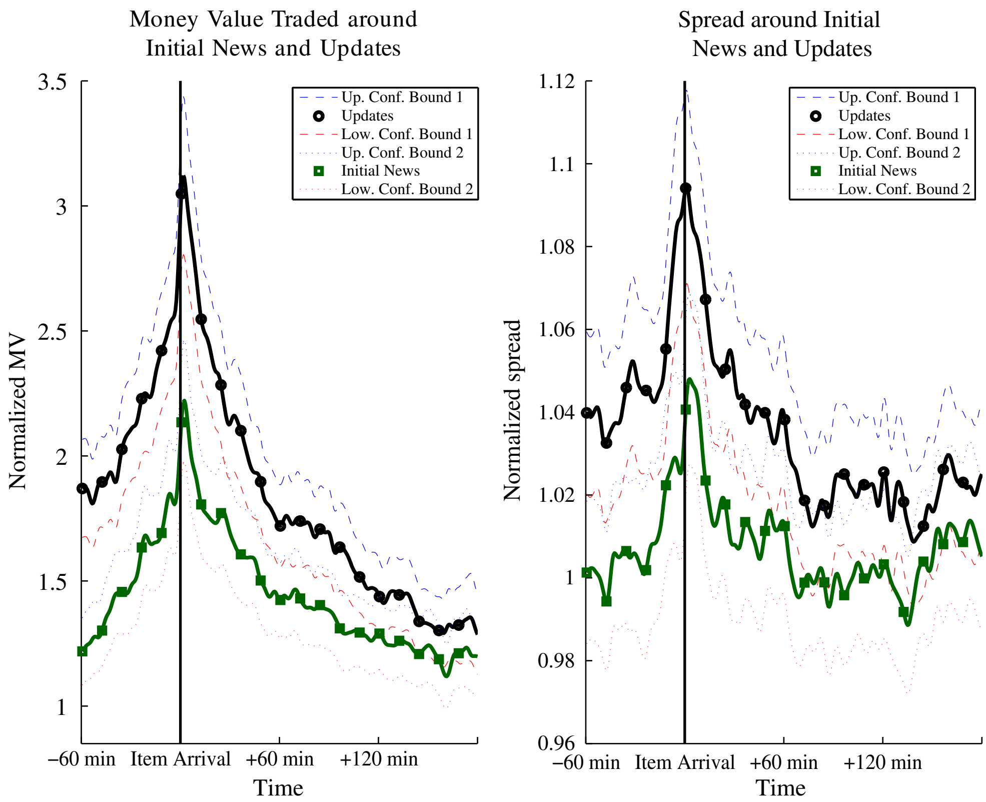

Fig. 6. Money value and spread around initial news and updates. Smoothed via kernel regression.

3.2. Trading profitability

To test for the profitability of trading on news items we employ an event study framework as outlined in Campbell et al. (1997). As a model for 'normal' returns we assume the market model

$$R_{it} = \alpha_i + \beta_i R_{mt} + \gamma_i R_{i,t-1} + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \left(0, \sigma_i^2\right), \tag{1}$$

where t denotes the underlying (20 s) intervals, R_{mt} is the market return, computed as the return of the FTSE 100 index, and R_{it} is the return for stock i. To capture return dynamics on high frequencies we also include lagged returns. Model (1) is estimated based on the complete 20-second return time series *without* including the event windows. Using the resulting parameter estimates, we compute the abnormal returns $\widehat{AR}_{it} := R_{it} - \hat{\alpha}_i - \hat{\beta}_i R_{mt} - \hat{\gamma}_i R_{i,t-1}$ during the event windows. Let $\widehat{\mathbf{AR}}_i^k$ denote the $((T_1+T_2+1)\times 1)$ vector of abnormal returns for event k of stock i computed between time points I_{-T_1} and I_{T_2} in Fig. 2. Let γ_j be a $(j \times 1)$ vector consisting of j ones, $1 \le j \le T_2 + T_1 + 1$. Then, we define the cumulated abnormal return for interval j after the I_{-T_1} interval as

$$\widehat{CAR}_{ij}^k := \gamma_j' \widehat{\mathbf{AR}}_i^k. \tag{2}$$

Averaging $\widehat{CAR}_{ij}^k$ yields

$$\overline{\widehat{CAR}}_j = \frac{1}{n}\left(\sum_i \sum_k \widehat{CAR}_{ij}^k\right), \tag{3}$$

where n is the total number of events over all stocks. Assuming (asymptotic) normality, 95% confidence intervals are computed as two times the standard deviation of the estimates $\overline{\widehat{CAR}}_j$.

Fig. 7 shows the averaged cumulated abnormal returns (ACAR) $\overline{\widehat{CAR}}$ around relevant news. Starting 90 min *before* the disclosure we observe significantly positive (negative) cumulated abnormal returns as reactions to positive (negative) news items. The news engine obviously allows to establish a significant relationship between a stories' sentiment and the corresponding sign of price trends. However, we observe already significant price movements prior to news releases but only little return reactions thereafter. Though private pre-release information might be present, we conjecture that the availability of other sources of information and an induced clustering of news items is mainly responsible for pre-announcement effects. Fig. 8 depicts the ACARs explicitly

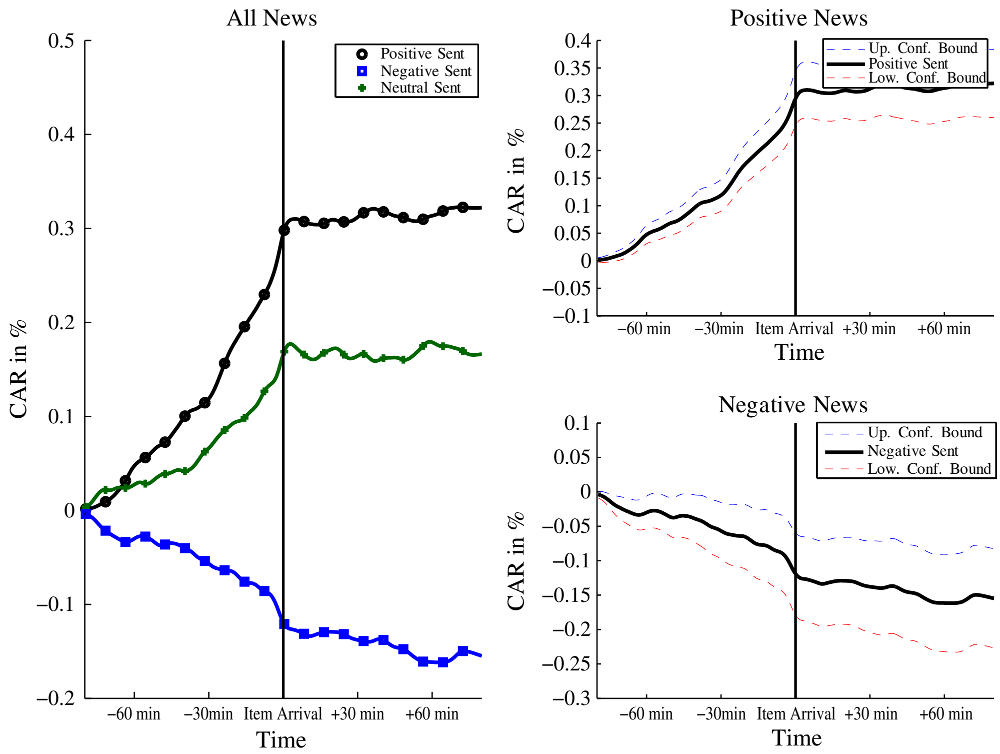

Fig. 7. Cumulated abnormal returns around relevant positive, negative and neutral news. Smoothed via kernel regression.

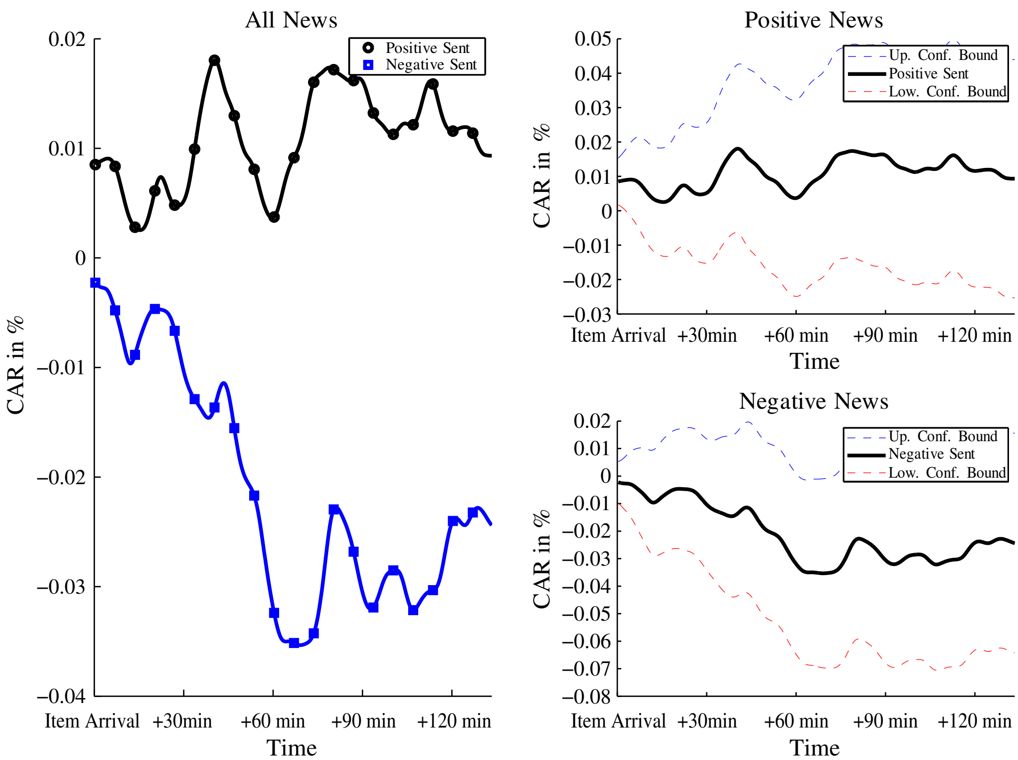

Fig. 8. Cumulated abnormal returns after relevant positive and negative news. Smoothed via kernel regression.

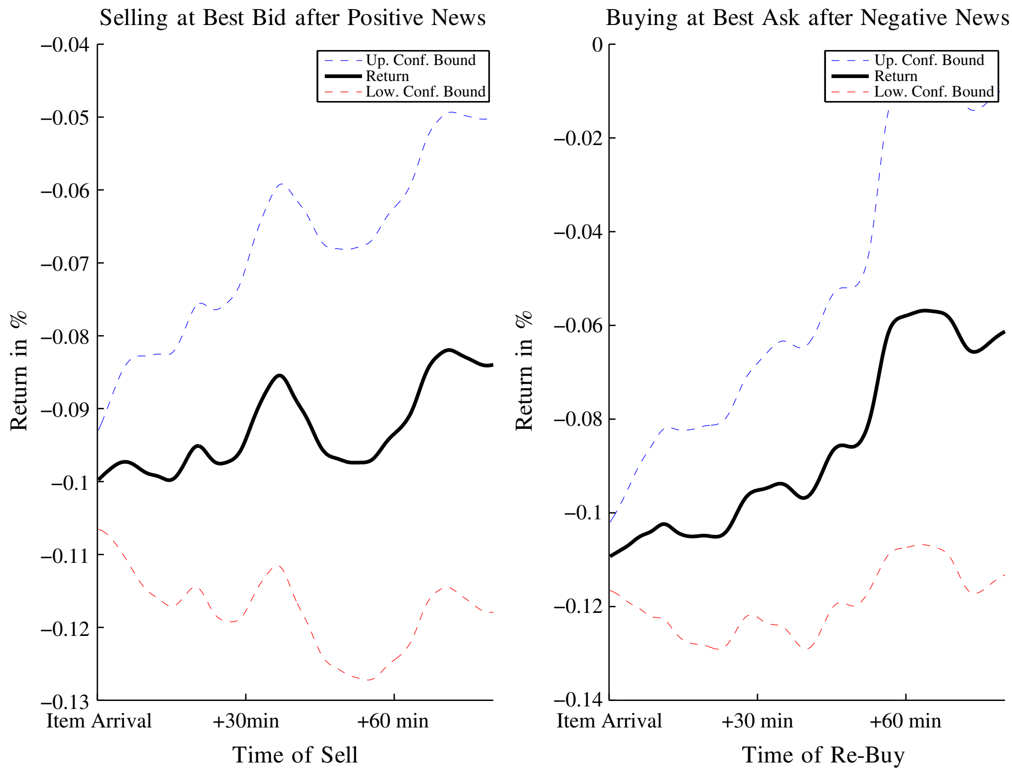

Fig. 9. Profitability of trading on the sentiments. Smoothed via kernel regression.

starting at news disclosure. We observe that sentiment indicators of news items have some predictability for future price movements. Nevertheless, we find the abnormal returns to be mostly insignificant.

In order to provide more specific evidence on trading profitability, we test a stylized trading strategy based on the sentiment information. Following a positive news item returns are computed by buying at the best ask at the item arrival and selling later at the bid. Conversely, after negative news, the asset is sold at the best bid and re-bought later at the ask. As shown by Fig. 9, we observe that the corresponding returns increase with the underlying horizon but are generally negative. This result shows that the abnormal returns of maximally 3.5 basis points in case of negative news (cf. Fig. 8) are too low to overcompensate increased bid-ask spreads around news and to provide economic gains of the underlying trading strategies. Still, the fact that return trends are positive (though never crossing the zero line), might be exploited via algorithmic trading strategies.

4. Market dynamics around news items

4.1. Econometric methodology

The unconditional analysis of the previous section provides strong indications for information-driven market reactions to news disclosures. However, we observe significant autocorrelations as well as cross-correlations between the variables. Fig. 10 shows corresponding autocorrelation functions.[3] The autocorrelation functions reveal a high persistence of the individual processes. Geweke and Porter-Hudak (1983) estimates of the fractional integration parameter (not shown in the paper) indicate that some series exhibit long range dependence and are overall covariance stationary. In order to avoid spurious results, the dependencies and interdependencies have to be explicitly taken into account. We suggest a six-dimensional model for the realized variance, the money value traded, the bid–ask spread, market depth, average trade size and absolute trade imbalance. As high-frequency volatility and liquidity variables are only weakly related to (signed) returns we refrain from including the latter in the model. Furthermore, in a separate analysis we find that the signed trade imbalances do not react to signed news (results given in Fig. 14, Appendix A). Nevertheless, to capture the order flow we consider the absolute trade imbalance instead. Accordingly, the vector of endogenous variables consists of the money value traded, the volatility, the absolute trade imbalance, the bid-ask spread, the market depth and the average trade size.

[3] Cross-correlations are given in Fig. 15 and Fig. 16 in Appendix A.

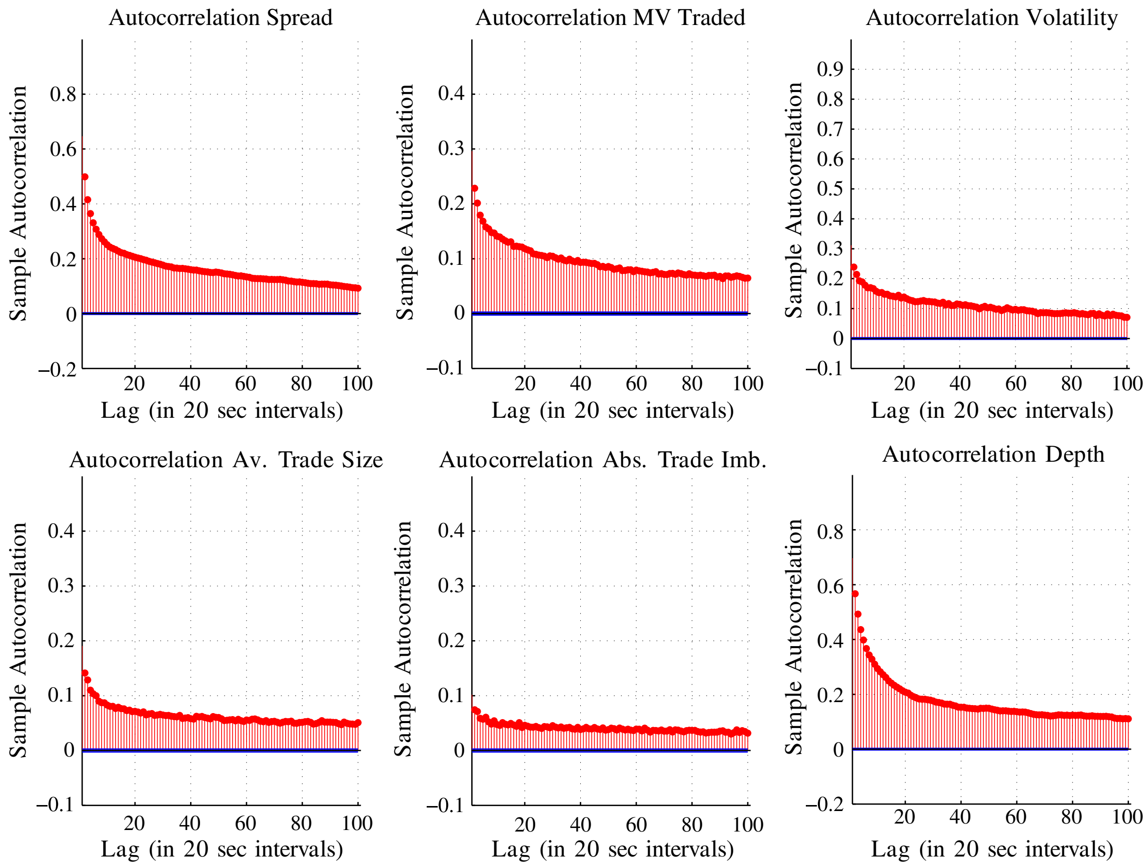

Fig. 10. Typical autocorrelation functions for the variables of interest (case: AV.L).

Since even for liquid stocks there is not necessarily a transaction in every 20 s interval we observe a non-trivial fraction of zero observations for money value traded and realized volatility. In particular, there are no trades in 46% of all 20 second intervals on average. To capture this finding, we suggest explicitly differentiating between the cases of trading, $y_{1t}>0$, and no trading, $y_{1t}=0$, in interval t. Correspondingly, the log likelihood function is given by

$$\ln \mathcal{L}(\mathbf{y};\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{\theta}_3) = \sum_{t=1}^{T} \{\ln f(\mathbf{y}_t|y_{1t}>0;\boldsymbol{\theta}_1) + \ln \mathrm{P}(y_{1t}>0;\boldsymbol{\theta}_2)\}\cdot \mathbb{1}(y_{1t}>0)$$
$$+\sum_{t=1}^{T} \{\ln \mathrm{P}(y_{1t}=0;\boldsymbol{\theta}_2) + \ln f(\mathbf{y}_t|y_{1t}=0;\boldsymbol{\theta}_3)\}\cdot \mathbb{1}(y_{1t}=0),$$

where $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$ denote corresponding parameter sets.

As long as the parameter sets $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$ are disjoint, the likelihood components can be maximized separately. Since $f(\mathbf{y}_t|y_{1t}=0;\boldsymbol{\theta}_3)$ is not in the core of our interest, we leave it unspecified. In addition, we refrain from explicitly modeling the long range dependence in some individual time series since this is infeasible in our case of 530,000 observations per variable and stock. To parameterize $f(\mathbf{y}_t|y_{1t}>0;\boldsymbol{\theta}_1)$, we suggest a VAR specification given by

$$\mathbf{y}_t|y_{1t}>0 = \mathbf{c} + \sum_{i=1}^{p} (\boldsymbol{\Gamma}_i\mathbf{y}_{t-i} + \boldsymbol{\Psi}_i Z_{t-i}) + \boldsymbol{\Xi}\cdot\mathbf{D}_t + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim N(\mathbf{0},\boldsymbol{\Omega}), \tag{4}$$

where $\boldsymbol{\Gamma}_i$ and $\boldsymbol{\Xi}$ denote (6×6) and $(6\times(p_1+p_2+1))$ coefficient matrices, where $p_1>0$ and $p_2>0$ are integers. Lags of the dummy $Z_t := \mathbb{1}_{(y_{1t}=0)}$ capture previous periods of non-trading with corresponding (6×1) coefficient vectors $\boldsymbol{\Psi}_i$. To capture the impact of news we define the dummy variable d_t with value one in case of relevant news in t and zero otherwise. Then, $\mathbf{D}_t := (d_{t+p_1},\ldots,d_{t-p_2})'$ is a vector of time dummies indicating the arrival of *relevant* news and covering p_1 intervals before and p_2 intervals after news disclosures. Model (4) can be consistently (though not necessarily efficiently) estimated equation by equation using ordinary least squares.

Table 2
Average VAR results: dynamics ($\mathbf{y}|y_{1t}>0$) and dummies.

		Money value	Volatility	Abs. trade imb.	Depth	Spread	Av. trade Sz.	Probit $\mathbb{1}(y_{1t}>0)$
Dynamics								
	c	-0.666^{***}	-0.120^{***}	0.522^{***}	0.239^{***}	0.342^{***}	0.159^{***}	1.847^{***}
		(0.115)	(0.076)	(0.010)	(0.017)	(0.022)	(0.038)	(0.048)
Money value	mv_{t-1}	0.224^{***}	0.046^{***}	-0.003^{***}	-0.001^{*}	0.001^{***}	0.020^{***}	0.018^{***}
		(0.003)	(0.005)	(0.000)	(0.000)	(0.000)	(0.002)	(0.003)
	mv_{t-2}	0.130^{***}	0.011^{***}	-0.001^{***}	0.001^{*}	-0.000	0.008^{***}	0.004^{***}
		(0.002)	(0.002)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)
Volatility	$vola_{t-1}$	0.000	0.203^{***}	-0.002^{***}	0.000	0.004^{***}	0.008^{***}	0.011^{*}
		(0.001)	(0.004)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)
	$vola_{t-2}$	-0.000	0.095^{***}	-0.001^{***}	0.000	0.002^{***}	0.004^{***}	0.003
		(0.001)	(0.002)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Abs. trade imb.	ati_{t-1}	0.140^{***}	0.118^{***}	0.060^{***}	0.002^{*}	-0.008^{***}	-0.053^{***}	0.222^{**}
		(0.017)	(0.016)	(0.001)	(0.002)	(0.002)	(0.007)	(0.005)
	ati_{t-2}	0.079^{**}	0.012	0.043^{***}	-0.007	-0.003	-0.024	0.147^{***}
		(0.010)	(0.015)	(0.001)	(0.001)	(0.002)	(0.005)	(0.004)
Depth	dp_{t-1}	0.594^{***}	-0.302^{***}	0.008^{***}	0.445^{***}	-0.025^{***}	0.398^{***}	0.008^{***}
		(0.025)	(0.024)	(0.001)	(0.010)	(0.001)	(0.017	(0.002)
	dp_{t-2}	-0.028	0.008	-0.001	0.091^{***}	-0.001	-0.002	0.008^{***}
		(0.007)	(0.008)	(0.000)	(0.002)	(0.001)	(0.006)	(0.001)
Spread	spr_{t-1}	-0.359^{***}	0.647^{***}	0.006^{***}	-0.097^{***}	0.297^{***}	-0.105^{***}	-0.160^{***}
		(0.016)	(0.071)	(0.001)	(0.006)	(0.009)	(0.007)	(0.005)
	spr_{t-2}	0.053^{**}	0.044^{*}	-0.004^{***}	-0.012^{***}	0.070^{***}	0.002	0.028^{***}
		(0.006)	(0.019)	(0.001)	(0.001)	(0.004)	(0.003)	(0.002)
Av. trade sz.	ats_{t-1}	-0.030^{***}	0.034^{***}	-0.000	-0.007^{***}	0.001	0.130^{***}	0.005^{***}
		(0.003	(0.004)	(0.000)	(0.001)	(0.000)	(0.003)	(0.002)
	ats_{t-2}	-0.018^{**}	0.017^{***}	0.001	-0.000	0.000	0.094^{***}	0.005^{***}
		(0.003	(0.002)	(0.000)	(0.001)	(0.000)	(0.003)	(0.001)
Dummies								
Dummy leads	d_{t+2}	0.894	1.023	-0.013	0.018	0.004	0.211	0.030
		(0.329)	(0.354)	(0.005)	(0.012)	(0.013)	(0.079)	(0.015)
	d_{t+1}	0.485^{*}	0.484^{*}	-0.009	0.002	-0.024	0.101	0.065
		(0.145)	(0.182)	(0.005)	(0.007)	(0.011)	(0.055)	(0.018)
Item dummy	d_t	1.036^{***}	1.332^{***}	-0.015	0.022	0.043	0.163	0.110
		(0.224)	(0.308)	(0.004)	(0.012)	(0.014)	(0.042)	(0.019)
Dummy lags	d_{t-1}	1.244^{***}	1.470^{***}	-0.022	-0.011	0.035	0.201	-0.081
		(0.236)	(0.482)	(0.004)	(0.014)	(0.010)	(0.081)	(0.016)
	d_{t-2}	0.891^{***}	0.925^{***}	-0.009	-0.003	0.020	0.137	0.048
		(0.178)	(0.330)	(0.006)	(0.012)	(0.010)	(0.048)	(0.015)
	d_{t-3}	0.588	0.909	-0.009	-0.015	-0.004	0.166	0.031
		(0.187)	(0.244)	(0.005)	(0.013)	(0.010)	(0.042)	(0.018)
	d_{t-4}	0.434	0.413	-0.013	0.020	0.005	0.107	0.057
		(0.145)	(0.145)	(0.005)	(0.013)	(0.010)	(0.057)	(0.019)
	d_{t-5}	0.307	0.549	-0.013	0.022	-0.000	0.076	0.024
		(0.179)	(0.189)	(0.004)	(0.013)	(0.011)	(0.040)	(0.018)

Note: The first six columns show OLS estimation results of system (4). The last column shows the ML estimation results of the corresponding probit model (5). The table gives estimates of the dynamics and estimates for the news dummies. Reported coefficients are averages of the estimates for each individual stock. Significance is reported based on average t-statistics. (Cross-sectional) standard errors of the averaged coefficients are given in parentheses below. *** denotes significance of the average coefficient estimates at the 1% level, ** at the 5% level, and * at the 10% level.

For the money value equation, the conditional probabilities for the occurrence of zero observations (i.e., no trading) in period t, $\mathrm{P}(y_{1t}=0;\boldsymbol{\theta}_2)$, are parameterized in terms of a probit specification. Let $\mathbf{x}_t$ contain all right-hand side variables of Eq. (4), i.e., $\mathbf{x}_t' := \left[1\,\mathbf{y}_{t-1}' \ldots \mathbf{y}_{t-p}'\, Z_{t-1}' \ldots Z_{t-p}' \mathbf{D}_t\right]$. Assuming a normally distributed latent process $y_{1t}^{*} \sim N(\mathbf{x}_t'\boldsymbol{\theta}_2, 1)$ underlying the trading "decision", we have

$$\mathrm{P}(y_{1t}^{*} > 0) = \Phi(\mathbf{x}_t'\boldsymbol{\theta}_2), \qquad \text{if } y_{1t}^{*} > 0 \Leftrightarrow y_{1t} > 0, \tag{5}$$

$$\mathrm{P}(y_{1t}^{*} \leq 0) = 1 - \Phi(\mathbf{x}_t'\boldsymbol{\theta}_2), \qquad \text{if } y_{1t}^{*} \leq 0 \Leftrightarrow y_{1t} = 0, \tag{6}$$

for the binary decision $y_{1t}>0$ vs. $y_{1t}=0$. The probit model is straightforwardly estimated by maximum likelihood.

The model is applied to each stock in our sample. Depending on the number of underlying trading days, the individual time series for the 39 stocks contain up to 533,000 observations. In order to obtain equal lag structures in all equations which eases cross-sectional comparisons and the computation of cross-sectional averages, we choose a universal lag length of 10 for all stocks. This lag length is sufficiently close to the individually optimal lag length according to the Bayes Information Criterion and does not

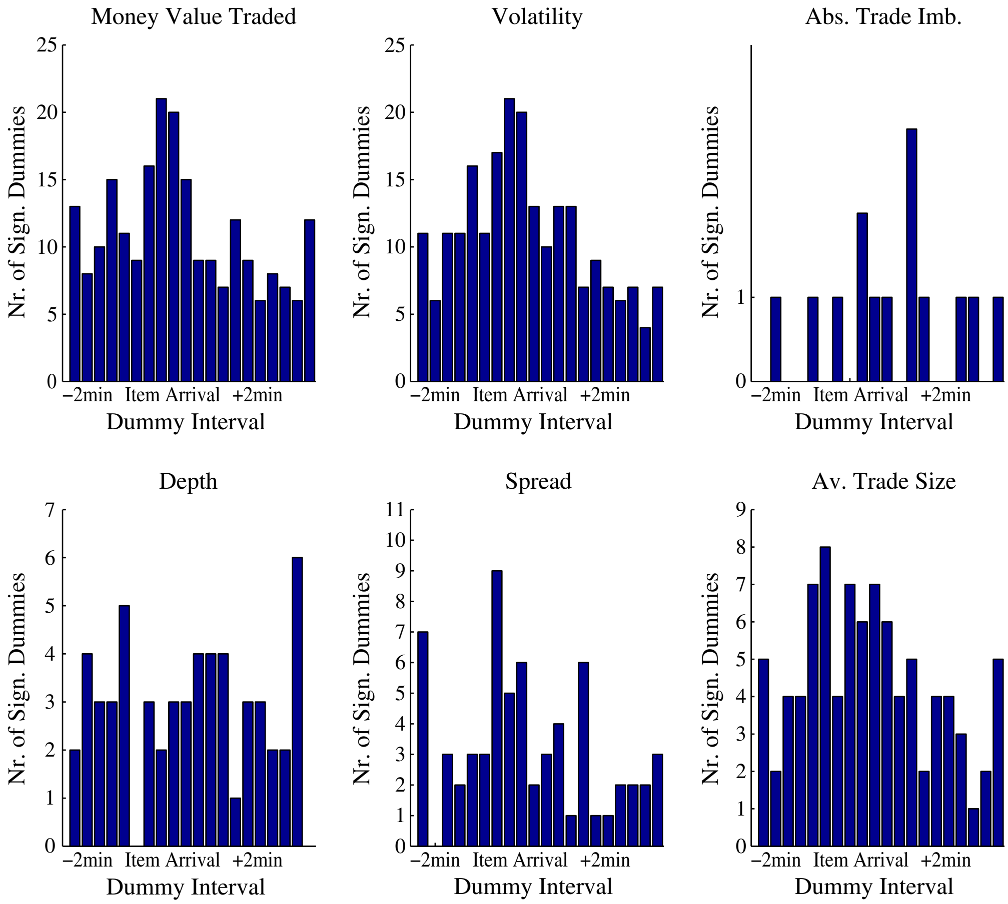

Fig. 11. Numbers of significant dummy variables in the intervals around the news disclosure. Dummies cover 7 20-second intervals before and 13 after the news arrival.

restrict the validity of the results discussed below. In the following we show the cross-sectional averages of point estimates and corresponding standard errors.

4.2. Estimation results

Table 2 reports averaged estimates of the VAR model which is augmented by dummies indicating relevant news. For sake of brevity, we do not show coefficients for lags of the dependent variables greater than two. Likewise, coefficient estimates for the dummies Z_t are not reported.[4] News dummies cover 40 s before the disclosure and 100 s thereafter.

Analyzing the dynamics of volatility and liquidity, we can summarize the following main results: First, all variables reveal significantly positive own dynamics. This is strongly expected given the underlying autocorrelations reported above. Second, we observe a significantly positive relationship between money value traded and volatility.

Hence, volatility and trading activity are closely dependent not only on a daily level as, e.g., shown by Clark (1973) and Tauchen and Pitts (1983), but obviously also on a high-frequency level, confirming, e.g., Hautsch (2008). Third, bid-ask spreads increase if past trading periods reflect rising liquidity demand and volatility. This causality is well confirmed by asymmetric information based market microstructure theory (e.g., Easley and O'Hara (1992)) where increased trading activity is an indicator for the existence of information and thus increased risks due to adverse selection. Our findings show that such situations are typically also characterized by increased trade sizes. Conversely, liquidity demand is reduced as a response to increased trading costs as induced by higher bid–ask spreads and reduced market depth.

Fourth, in contrast to the unconditional analysis in Section 3, significant effects induced by news items are only identifiable for volatility and cumulated trading volumes but not for spreads, absolute trade imbalances, average trade sizes and market depths. In particular, the insignificant spread dummies contradict corresponding results for earnings announcements. These results suggest that the (unconditional) reactions of these variables during news arrival periods as reported in Section 3 are mainly due to spillover effects arising from increased volatility and cumulated volumes but do not necessarily arise from news alone. Moreover,

[4] These results are available upon request from the authors.

due to the persistence in the market dynamics, news-induced effects and pre-release trading activity are carried over to subsequent periods. It is therefore not surprising that the direct impact of news as captured by the dummy variables dies out relatively quickly. These findings show that ignoring dynamics and interdependencies can cause spurious results.

Fifth, confirming the results of Section 3 we find significant effects only for relevant news. Indeed, filtering out noise and structuring news according to their relevance is even more important in a dynamic setting than in an unconditional framework.

The estimation results for the probit model widely confirm those for the VAR specification. However, the fact that all news dummy variables are insignificant indicates that the probability for the occurrence of a trade in a 20 s interval is not significantly driven by news arrivals.

The averaged estimates capture the major features common to all assets, but most stocks still reveal idiosyncratic responses to news. Even though, for instance, the *average* spread reaction is insignificant, we still observe significant individual spread responses for 19 out of 39 stocks in the sample. Fig. 11 shows that the significant (positive) dummies for most stocks center around the item arrival interval. Accordingly, there is evidence for news-implied reactions in spreads, depth and average trade sizes which are, however, diffuse across the stock universe. Stock-specific effects for the money value and volatility are much more in line with the average results as we detect significant reactions after news arrivals for all but three stocks.

4.3. Impulse response analysis

To provide more insights into news-implied market responses in a dynamic system, we perform an impulse response analysis. Here, a 'news shock' is defined as a change in the corresponding news dummies. As the arrival of news generally stimulates trading activity, it is sufficient to conduct the analysis under the assumption that there is trading activity throughout post-release periods, i.e., $\mathbf{y}_j | y_{1j} > 0$ for all $j = t, \ldots, t+s$.

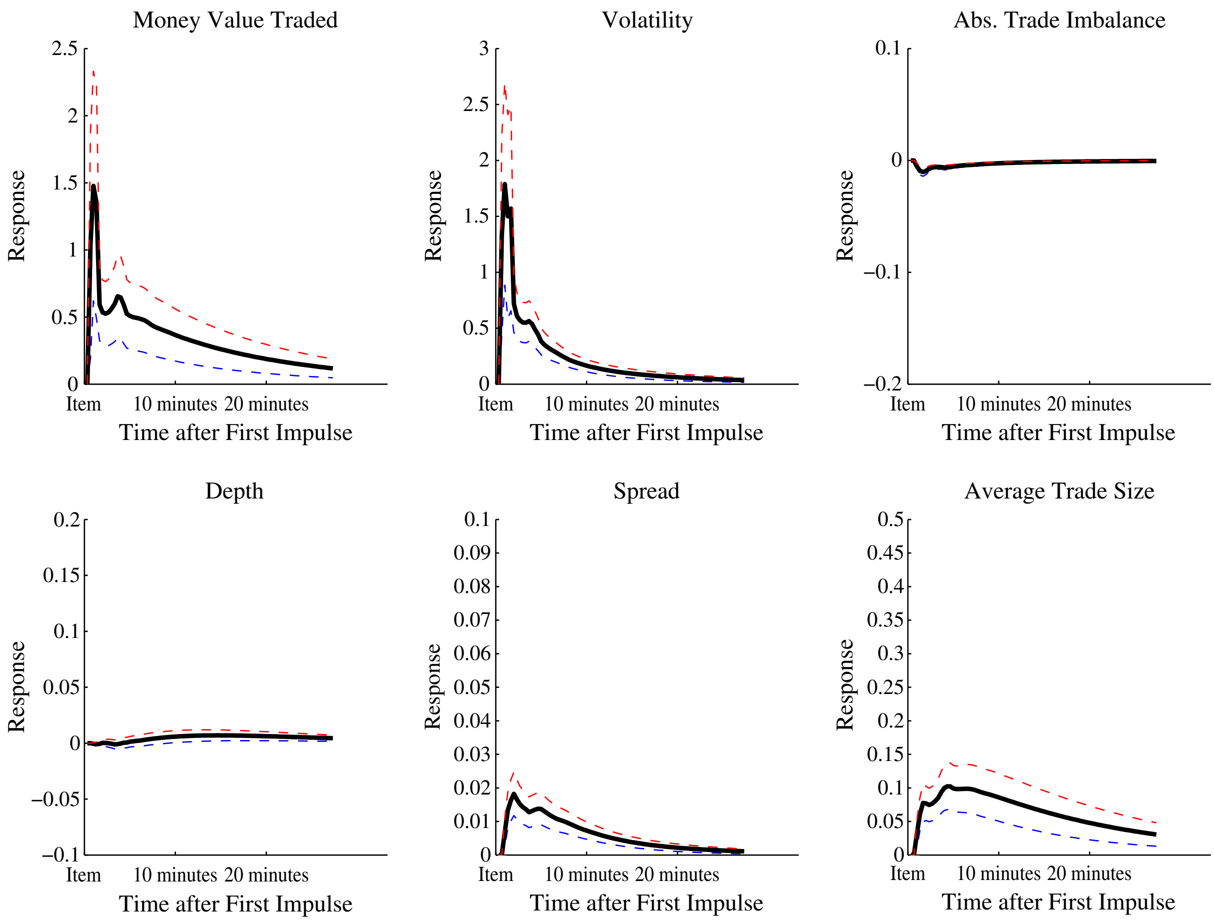

Fig. 12. Response analysis of a change in the news dummies for highly relevant news (95% confidence intervals as dotted lines).

Accordingly, the response after s periods to a news arrival in t is computed as

$$\Delta_s(\boldsymbol{\theta}_1) \ := \mathbb{E}[\mathbf{y}_{t+s}|\Omega_{t-1}, d_t = 1; \boldsymbol{\theta}_1] - \underbrace{\mathbb{E}[\mathbf{y}_{t+s}|\Omega_{t-1}, d_t = 0; \boldsymbol{\theta}_1]}_{(*)}, \tag{7}$$

where Ω_{t-1} represents the history of the multivariate process at t and the second term $(*)$ removes the effect of constants and initial values on the response function. Let $p_1 = 0, p_2 > 0$ and $\hat{\Xi}_{\cdot i}$ denote the i-th column of $\hat{\Xi}$. Coefficients in the second to p_2-th columns of $\hat{\Xi}$ that are not significantly different from zero at the 5% level are assumed to be zero throughout. Initially we have

$$\begin{aligned}\Delta_0 &= \mathbb{E}[\mathbf{y}_t|\Omega_{t-1}, d_t = 1; \boldsymbol{\theta}_1] - \mathbb{E}[\mathbf{y}_t|\Omega_{t-1}, d_t = 0; \boldsymbol{\theta}_1] \\ &= \hat{\mathbf{c}} + \sum_{i=1}^{p} \left(\hat{\boldsymbol{\Gamma}}_i \mathbf{y}_{t-i} + \hat{\boldsymbol{\Psi}}_i Z_{t-i}\right) + \hat{\Xi}_{\cdot 1} - \left(\hat{\mathbf{c}} + \sum_{i=1}^{p} \left(\hat{\boldsymbol{\Gamma}}_i \mathbf{y}_{t-i} + \hat{\boldsymbol{\Psi}}_i Z_{t-i}\right)\right) = \hat{\Xi}_{\cdot 1}.\end{aligned}$$

Since the initial conditions, constants and Z_t cancel out, the responses in $t+s, s = 1, 2, ...,$ to the dummy impulse in t are given as

$$\Delta_1 = \hat{\boldsymbol{\Gamma}}_1 \Delta_0 + \hat{\Xi}_{\cdot 2}, \qquad \Delta_2 = \hat{\boldsymbol{\Gamma}}_1 \Delta_1 + \hat{\boldsymbol{\Gamma}}_2 \Delta_0 + \hat{\Xi}_{\cdot 3}, \ldots$$

Standard errors of the response function are derived using the delta method. Accordingly, Δ_s is asymptotically distributed as

$$\Delta_s\left(\hat{\boldsymbol{\theta}}_1\right) \xrightarrow{d} N\left(\Delta_s(\boldsymbol{\theta}_1), (1/T)\mathbf{G}_s\left(\boldsymbol{\Omega} \otimes \mathbf{Q}^{-1}\right)\mathbf{G}_s'\right),$$

where $\mathbf{Q} = \mathbb{E}[\mathbf{x}_t \mathbf{x}_t']$ and $\mathbf{G}_s = \dfrac{\partial \Delta_s(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_1'}$. Estimates for $\boldsymbol{\Omega}$ and $\mathbf{Q}$ are readily available from the VAR estimates. Following Hamilton (1994), we construct the columns of estimates of $\mathbf{G}_s$ based on finite differences according to

$$\frac{\partial \Delta_s\left(\hat{\boldsymbol{\theta}}_1\right)}{\partial \boldsymbol{\theta}_{1i}} \approx \frac{\Delta_s\left(\hat{\boldsymbol{\theta}}_1 + \mathbf{e}_i h\right) - \Delta_s\left(\hat{\boldsymbol{\theta}}_1\right)}{h},$$

where h is some small number, θ_{1i} denotes the i-th element of $\boldsymbol{\theta}_1$ and e_i is the i-th unity vector.

Fig. 12 shows the impulse response to news-induced dummy variable changes based on the averaged VAR estimates. The depicted reaction to relevant news mimics the unconditional market responses of volatility, money value traded, average trade sizes and bid–ask spread very well (cf. Figs. 3–5). The percentage increase of volatility and money value traded over the pre-news period is roughly 150% (given the mean one for each series) and thus corresponds to the unconditional effects. Again it turns out that responses in bid–ask spreads and trade sizes are induced through dynamic spillovers from news effects in volatility and cumulative volumes.

Overall, we conclude that the dynamic analysis strongly confirms the unconditional effects shown above. Obviously, volatility and trading volume are most sensitive to news arrival. Reactions in bid-ask spreads and average trade sizes are rather idiosyncratic and due to spillovers. In order to check the robustness of our results, we have estimated several alternative specifications, in particular (i) a simple VAR model based on 20 s aggregates (without explicitly accounting for zero observations), and (ii) the corner-solution model by Cragg (1971) for the conditional density based on 20 s aggregates. For sake of brevity we refrain from reporting the corresponding estimates in the paper. However, it turns out that our findings are qualitatively quite stable across the individual specifications.

5. Conclusions

Recording and analyzing the overall news flow for a specific asset is challenging since the amount of news, the number of news sources and the speed of information dissemination is rapidly increasing over time. Due to the huge amount of information permanently published in all modern media, news are overlaid by substantial noise caused by irrelevant information. These effects make it difficult to identify significant links between high-frequency trading activity and the intraday news flow. As previous studies predominantly focus only on scheduled and homogenous types of news (typically earnings announcements), it is virtually unknown whether intraday trading activity, volatility and liquidity can be systematically linked to an intraday flow of news items other than regularly announced earnings figures.

To reduce the impact of noise, this study is the first one making use of unique data provided by an automated news analytics tool of the Reuters Company. Designed for use in algorithmic trading applications and employing linguistic pattern recognition techniques, these novel news data allow us to disentangle relevant news from irrelevant ones and to identify the sign and the novelty of news items. Using this news engine our study explores the impact of news items on high-frequency returns, trading volume, volatility, depth and bid–ask spreads for a cross-section of stocks traded at the London Stock Exchange (LSE).

Analyzing the unconditional and conditional effects of news items on intraday trading activity we can summarize the following results. First, we identify significant unconditional market responses in volatility, money value traded, average trade sizes and bid–ask spreads. Given the fact that earnings announcements are explicitly discarded from the analysis these findings are remarkable and indicate that the news engine successfully filters the news flow. However, it turns out that only cumulative trading volumes and volatility are directly influenced by news releases, whereas reactions in bid–ask spreads and trade sizes are widely indirect and due to dynamic spillovers from volatility and volumes. Second, we confirm the usefulness of the machine-indicated relevance of news items. In fact, significant market responses to news are only observable for items which are identified as being relevant. Our results show that the classification is crucial to filter out noise and to identify significant relations between market activity and the news flow. Third, the news sentiment indicator has predictability for future price trends. However, significantly increased bid–ask spreads around public news arrivals render simple sentiment-based trading strategies rather unprofitable.

Overall, our study shows that news engines are able to successfully structure and categorize the intraday news flow. This allows to deeply investigate the question to which extent high-frequency market activity is driven by information. Moreover, our findings provide first evidence on the usefulness of news engines in financial practice. In specific, the distinct volatility and liquidity effects around news are potentially relevant parameters in algorithmic trading applications.

Appendix A. A note on the computation of standard errors of across-market averages

In the following we describe two ways of computing the mean reactions and their standard errors. The pooled average used in Section 3 is based on the model

$$X_i = \mu + \epsilon_i, \quad \epsilon_i \sim i.i.d.\ N\left(0, \sigma^2\right), \quad i = 1, \ldots, n, \tag{8}$$

where we have suppressed the I_j index for the respective interval around the news item. Inference is based on the pooled estimator for the mean, $\overline{X} = 1/n\sum_{i=1}^{n} X_i$, where 95% confidence intervals are given as $\overline{X} \pm 2 * \hat{\sigma}/\sqrt{n}$ with $\hat{\sigma}^2 = \mathbf{e}'\mathbf{e}/(n-1)$.

To account for the fact that the stocks have very different numbers of news items (see Table 1), we alternatively used group-specific means. Let n_s denote the number of news for stock s and let X_{sj} be the reaction of a certain (trading) variable of stock s to item j. For the average reaction of the n_n stocks with individual means, $\overline{X}_s = 1/n_s\sum_{j=1}^{n_s} X_{sj}$, we assume

$$\overline{X}_s = \mu + \epsilon_s, \qquad \epsilon_s \sim i.i.d.\ N\left(0, \sigma^2\right), \qquad s = 1, \ldots, n_n. \tag{9}$$

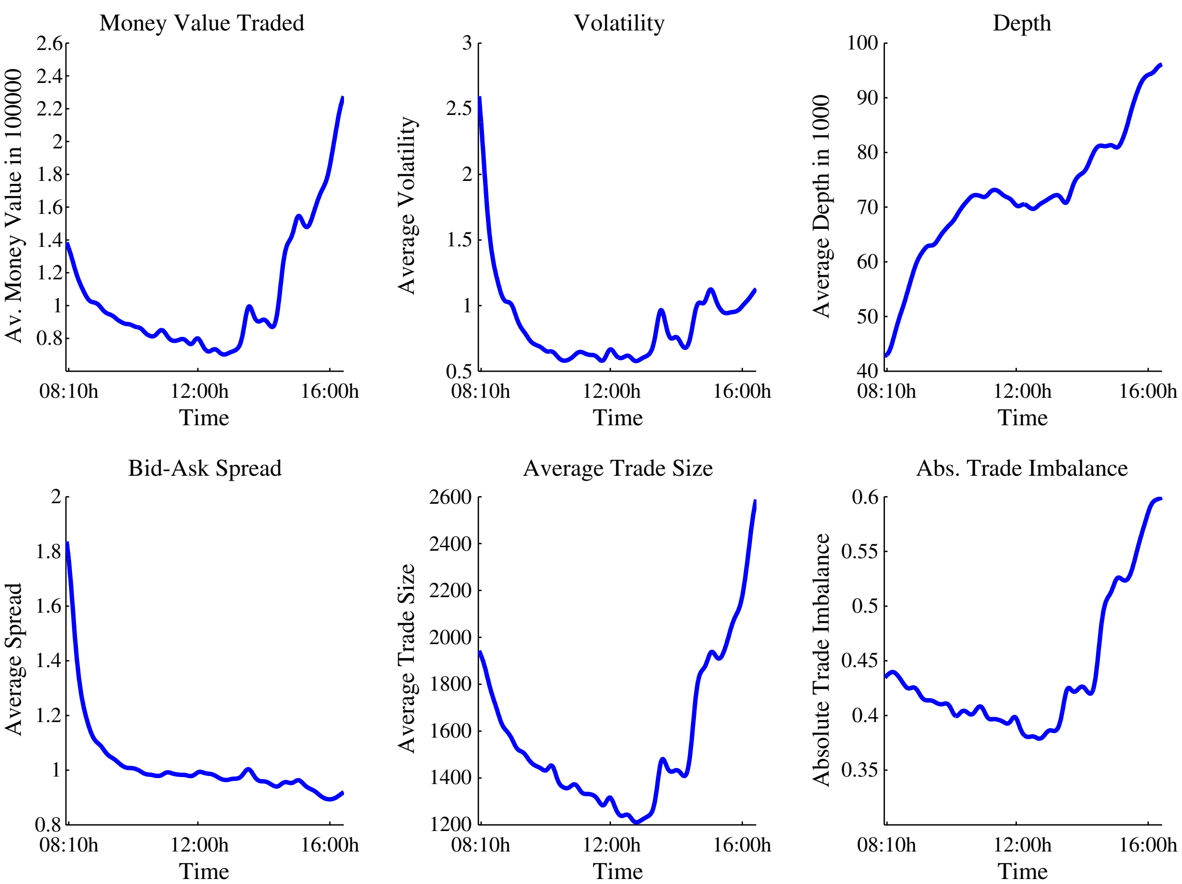

Fig. 13. Average intraday seasonality patterns. Smoothed via kernel regression.

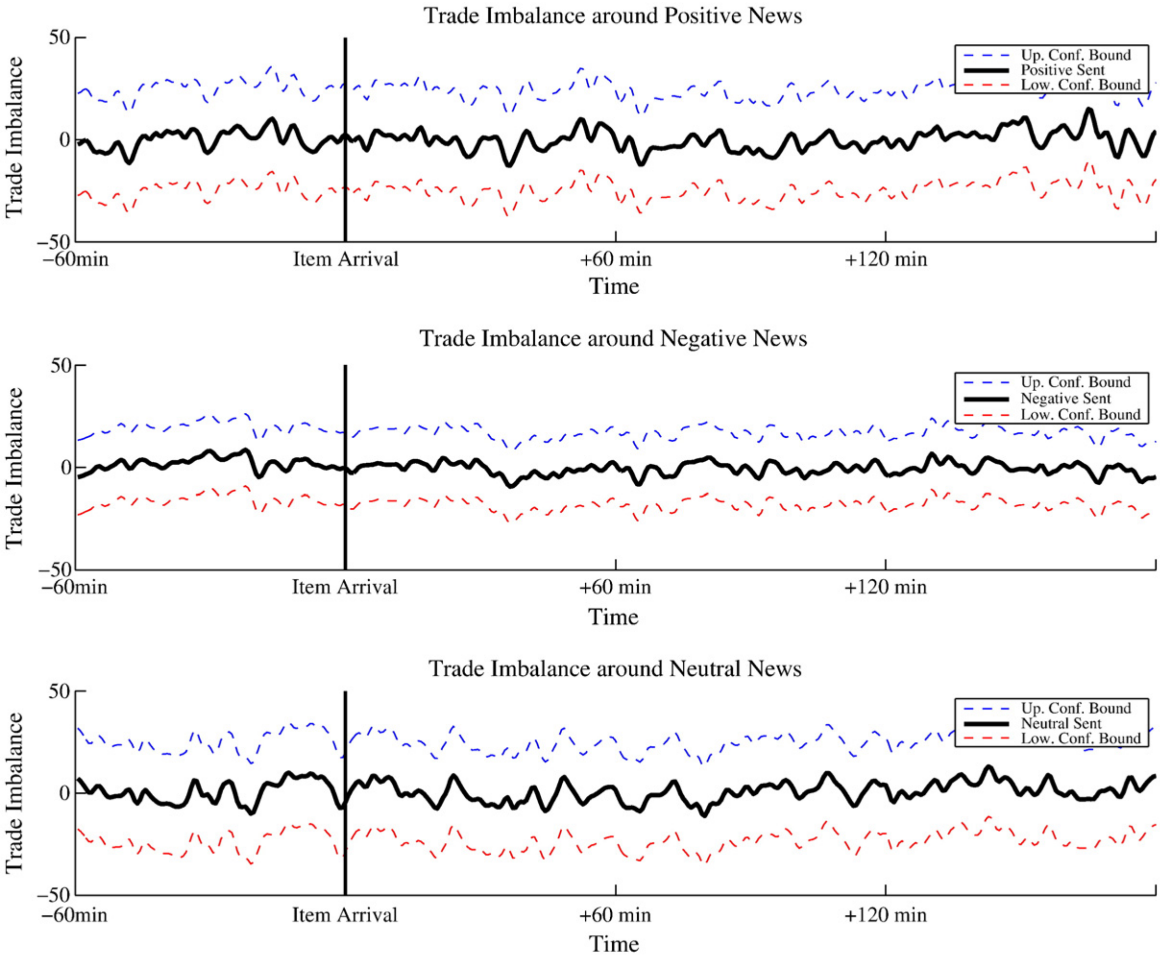

Fig. 14. Signed trade imbalance around positive, negative and neutral news. Smoothed via kernel regression.

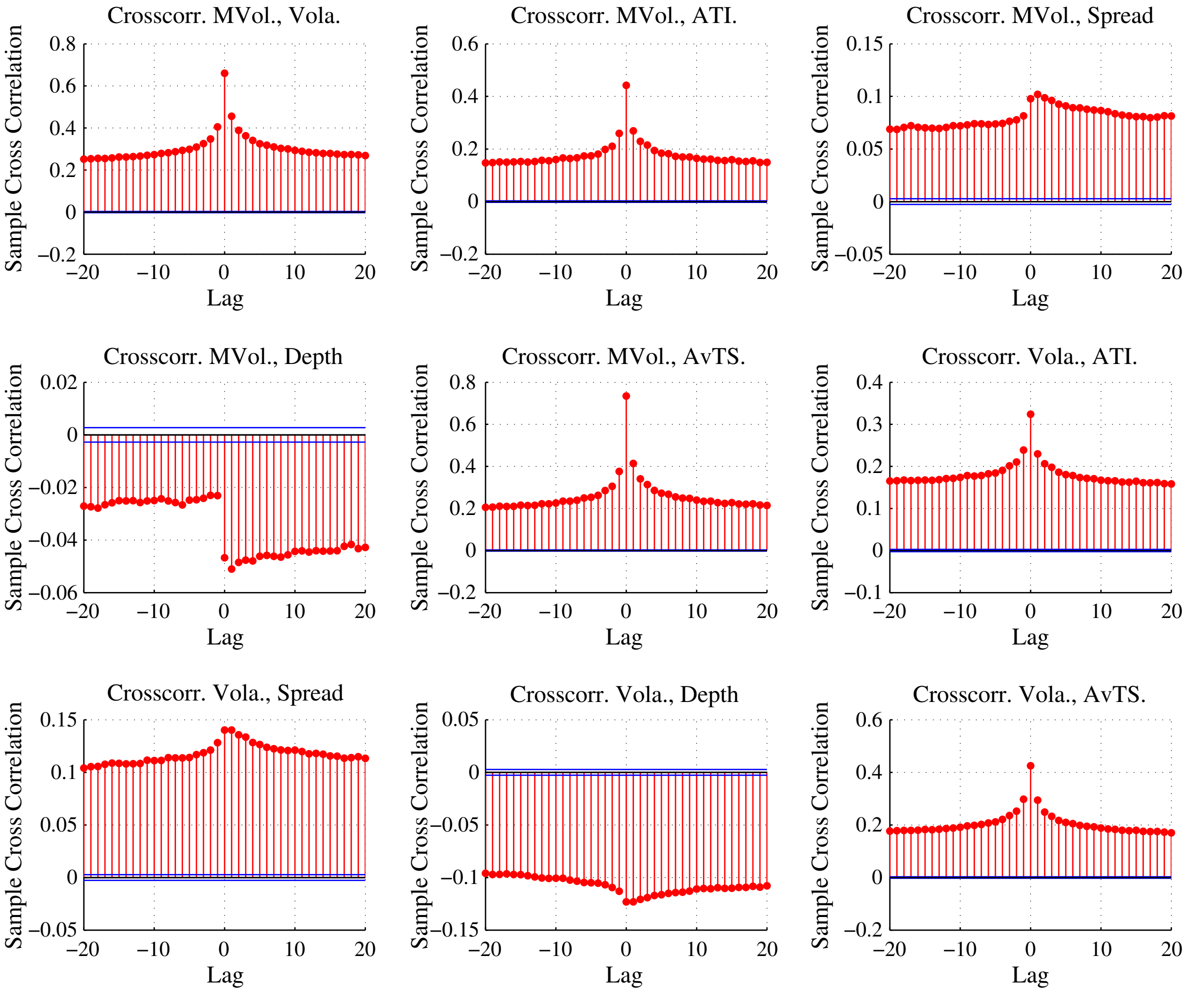

Fig. 15. Sample average of cross-correlations (I).

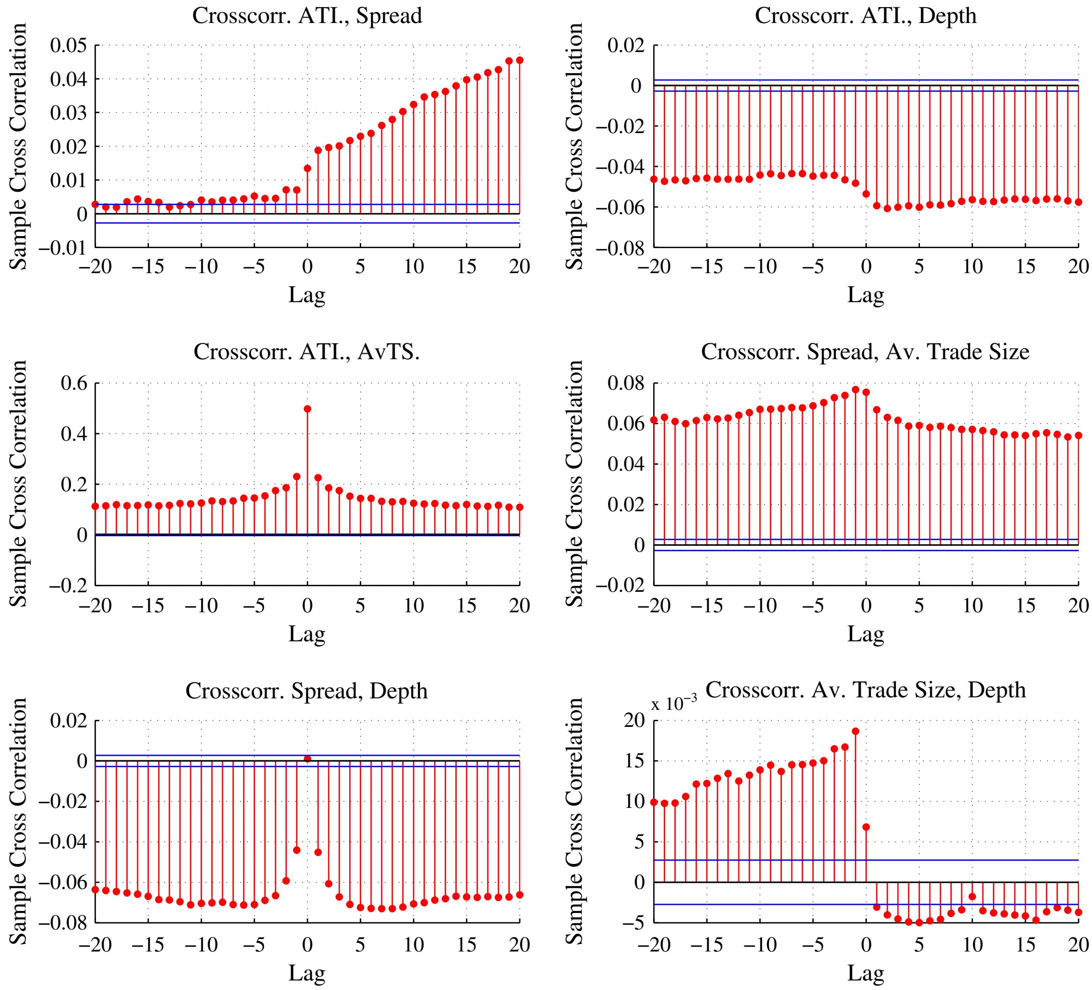

Fig. 16. Sample average of cross-correlations (II).

Then, inference is based on the estimator for the mean, $\overline{\overline{X}} = 1/n_n \sum_{s=1}^{n_n} \overline{X}_s$, where 95% confidence intervals are given as $\overline{\overline{X}} \pm 2 * \hat{\sigma}/\sqrt{n_n}$ with $\hat{\sigma}^2 = \mathbf{e}'\mathbf{e}/(n_n - 1)$.

Both approaches have their advantages. While the latter smoothes out the effect of a large number of news, it does not account for the within-group variation, which is captured by Eq. (8). Hence, confidence intervals are slightly more conservative using Eq. (9). Nevertheless, all results of Section 3 hold using both procedures. Plots of the group-means are available upon request from the authors.

References

Antweiler, W., Frank, M.Z., 2004. Is all that talk just noise? The information content of internet message boards. J. Finance 59, 1259–1293.
Barber, B.M., Odean, T., 2008. All that glitters: the effect of attention and news on the buying behavior of individual and institutional investors. Rev. Financ. Stud. 21, 785–818.
Berry, T.D., Howe, K.M., 1994. Public information arrival. J. Finance 49, 1331–1346.
Blume, L., Easley, D., O'Hara, M., 1994. Market statistics and technical analysis. J. Finance 49, 153–181.
Busse, J., Green, T.C., 2002. Market efficiency in real time. J. Financ. Econ. 65, 415–437.
Campbell, J.Y., Lo, A.W., MacKinlay, A.C., 1997. The Econometrics of Financial Markets. Princeton University Press, Princeton, New Jersey.
Clark, P.K., 1973. A subordinate stochastic process model with finite variance for speculative prices. Econometrica 41, 135–155.
Cragg, J., 1971. Some statistical models for limited dependent variables with application to the demand for durable goods. Econometrica 39, 829–844.
Easley, D., O'Hara, M., 1987. Price, trade size and information in securities markets. J. Financ. Econ. 19, 69–90.
Easley, D., O'Hara, M., 1992. Time and the process of security price adjustment. J. Finance 47, 577–605.
Fleming, M.J., Remolona, E.M., 1999. Price formation and liquidity in the U.S. Treasury market: the response to public information. J. Bank. Finance 28, 1441–1467.
Geweke, J., Porter-Hudak, S., 1983. The estimation and application of long memory time series models. J. Time Ser. 4, 221–238.
Hamilton, J.D., 1994. Time Series Analysis. Princeton University Press, Princeton, New Jersey.
Harris, M., Raviv, A., 1993. Differences of opinion make a horse race. Rev. Financ. Stud. 6, 473–506.
Hautsch, N., 2008. Capturing common components in high frequency time series: a multivariate stochastic multiplicative error model. J. Econ. Dyn. Control 32, 3978–4009.
Kalev, P.S., Liu, W.S., Pham, P.K., Jarnecic, E., 2004. Public information arrival and volatility of intraday stock returns. J. Bank. Finance 28, 1441–1467.

Kandel, E., Pearson, N.D., 1995. Differential interpretation of public signals and trade in speculative markets. J. Polit. Econ. 103, 831–872.
Karpoff, J.M., 1986. A theory of trading volume. J. Finance 41, 1069–1087.
Kim, O., Verrecchia, R.E., 1991. Market reactions to anticipated announcements. J. Financ. Econ. 30, 273–309.
Kim, O., Verrecchia, R.E., 1994. Market liquidity and volume around earnings announcements. J. Acc. Econ. 17, 41–67.
Krinsky, I., Lee, J., 1996. Earnings announcements and the components of the bid-ask spread. J. Finance 51, 1523–1535.
Lee, C., Ready, M., 1991. Inferring trade direction from intraday data. J. Finance 46, 733–746.
Lee, C.M.C., Mucklow, B., Ready, M.J., 1993. Spread, depths and the impact of earnings information: an intraday analysis. Rev. Financ. Stud. 6, 345–374.
Mitchell, M.L., Mulherin, J.H., 1994. The impact of public information on the stock market. J. Finance 49, 923–950.
Ranaldo, A., 2008. Intraday market dynamics around public information disclosures. In: Lhabitant, F.-S., Gregoriou, G. (Eds.), Stock Market Liquidity. John Wiley and Sons, New Jersey, pp. 199–226. chap. 11.
Tauchen, G.E., Pitts, M., 1983. The price variability-volume relationship on speculative markets. Econometrica 51, 485–505.
Tetlock, P.C., 2007. Giving content to investor sentiment: the role of media in the stock market. J. Finance 62, 1139–1167.
Tetlock, P.C., 2010. Does public financial news resolve asymmetric information? Rev. Financ. Stud. 23, 3520–3557.
Tetlock, P.C., Macskassy, S.A., Saar-Tsechansky, M., 2008. More than words: quantifying language to measure firms' fundamentals. J. Finance 63, 1427–1467.