How does post-earnings announcement sentiment affect firms' dynamics? New evidence from causal machine learning

Francesco Audrino* Jonathan Chassot* Chen Huang† Michael Knaus*
Michael Lechner* Juan Pablo Ortega*,‡

December 2020

Abstract

We revisit the role played by sentiment extracted from news articles related to earnings announcements as a driver of firms' return, volatility, and trade volume dynamics. To this end we apply causal machine learning on the earnings announcements of a wide cross-section of US companies. This approach allows us to investigate firms' price and volume reactions to different types of post-earnings announcement sentiment (positive, negative, and mixed sentiments) under various underlying macroeconomic and aggregated investors' moods in a properly defined causal framework. Our empirical results support the presence of (i) investors' overconfidence and mispricing due to biased expectations; (ii) a leverage effect in sentiment where reactions are (on average) larger for negative sentiment; and (iii) investors' underreaction to news. Finally, we show that the difference in the average causal effects of the sentiment's types is larger when the general macroeconomic conditions are worse or the uncertainty in the global financial market is higher.

JEL classifications: C31; C58; G14, G41.

Keywords: Post-earnings announcement drift, volatility, and volume; Sentiment; Causal machine learning; Modified causal forest; Heterogeneity analysis.

*University of St. Gallen. Corresponding author. Address: Bodanstrasse 6, 9000 St.Gallen, Switzerland. Phone number: +41712242431. ORCID: 0000-0002-2038-9362. E-mail: francesco.audrino@unisg.ch

†Aarhus University

‡Centre National de la Recherche Scientifique (CNRS)

1 Introduction

Firms' earnings announcements are key moments in which investors receive real, correct information about the general companies' conditions and the performance achieved during the preceding period of time (usually one quarter). Investors then update their financial positions, correct potentially biased expectations and, depending on the type of news released and their previous perception of the companies' potential, keep or change their opinion and sentiment. As has already been widely reported in the literature, this release of new information has a huge impact on the dynamics of the companies' prices and volatilities and on the way investors trade the underlying stocks the days of and after the earnings calls. These reactions produce what are generally referred to as "anomalies," that is, empirical, observed facts that counter the expectations developed in standard theories that are based on the key assumptions of rational investors and efficient financial markets. In fact, abnormal returns, excess volatility and abnormal trading volumes are empirical stylized facts commonly associated with earnings announcements.

As a consequence, several models have been introduced in the literature that incorporate different possible investors' behaviors in an attempt to close the gap between theory and observed reality. Given the growing empirical evidence dating back to the early 1990s coming from behavioral finance showing that the stock market is driven by investors' psychology (see, among others, Daniel et al., 2002, for a literature review), all such models include as central variables investors' sentiment and attention and put forward different explanations (in isolation or combined with each other): Among others, the misattribution bias, which says that people make risky decisions depending on mood states (Johnson and Tversky, 1983); models assuming that risk-averse arbitrageurs encounter limits to arbitrage and cannot always keep the prices at their fundamental values when trading against the irrational trades of uninformed noise traders basing decisions on sentiment (Miller, 1977; Black, 1986; De Long et al., 1990; Stambaugh et al., 2012); underreaction and overreaction to news (Barberis et al., 1998; Hirshleifer et al., 2009); optimism and pessimism in the sense that investors are too optimistic about some stocks and too pessimistic about others (Engelberg et al., 2018); overconfidence (Miller, 1977); the limited attention hypothesis, which assumes that attention is a scarce and costly resource, and as such investors pay only fluctuating attention to stock-related information, reducing the speed at which the new information is properly interpreted and incorporated in stock prices (Kahneman, 1973; Peng, 2005; Peng and Xiong, 2006; Della Vigna and Pollet, 2009; Hirshleifer et al., 2009; Andrei and Hasler, 2015; Gargano and Rossi, 2018; Audrino et al., 2020); and investors' sentiment disagreement and divergence of opinions (Antweiler and Frank, 2004; Siganos et al., 2017; Giannini et al., 2019; Audrino et al., 2020; Cookson and Niessner, 2020).

In this empirical study we revisit the role played by the sentiment extracted from news articles related to the release of new information during firms' earnings calls in driving subsequent companies' price dynamics and investors' trading behaviors. In particular, we investigate whether different types of sentiment yield systematic various immediate and mid-term investors' reactions observed the day and the week after the earnings announcements. Moreover, we verify whether such reactions align well with the main underlying economic theories introduced in the previous literature.

In contrast to all previous studies investigating stock price dynamics around earnings announcements, we apply a recent methodology developed by the causal machine learning community, the Modified Causal Forest (MCF), to estimate the (causal) effects of interest. Such an approach allows us to control in a more systematic way for the information included in a possibly large number of confounding covariates and to clarify further both the direction and the size of the effects as well as to analyze heterogeneity in a systematic way. In greater detail, we apply a method based on causal forests proposed in Wager and Athey (2018) based on a collection of randomly generated causal trees with small final leaves. This approach is combined with the techniques introduced in Lechner (2018) that improve the splitting rule for the individual trees and provide methods to estimate heterogeneous effects for a limited number of discrete policy variables at low computational cost.

Using this methodology, our analysis focuses on the effect of different types of post-earnings announcement sentiment (categorized as positive, negative, and mixed) extracted from the news articles reporting and commenting on the earnings releases, individual firms' returns, volatilities, and trading volumes. Moreover, the flexibility of this strategy also allows performing the analysis on whether and how the estimated effects change with respect to different underlying macroeconomic conditions and aggregated investors' moods about the financial market as a whole.

In our analysis we consider a wide cross section of US companies belonging to the S&P 500 universe from September 10, 2003 to December 31, 2017. We obtained the data about the news articles' sentiment related to the firms' earnings releases through the RavenPack News Analytics database and the price and volume data through the TAQ database. In particular, we consider only earnings announcements that are released when the US market is closed (in the evening after the market closes or in the morning before market opening): This allows an exact identification of the beginning of the post-announcement period and performance of the analysis on a daily frequency basis. As we show in Section 2, this is not an overly restrictive assumption given that about 97% of the earnings calls happen when the market is closed.

The main results can be summarized as follows. First, we corroborate previous empirical evidence and find that the day and the week after the earnings announcements, the average potential returns due to the different news articles' sentiment are large (in magnitude): They go up to 21 basis points and 75 basis points on average over the cross-section of companies on a daily and weekly basis, respectively. Moreover, both average potential volatility and trading volume jump irrespectively of the sentiment's direction the day after the earnings announcements to levels that are more than 60% larger than usual. One week after the announcements, the average potential volatility and trading volume slowly revert back to the standard average values. These results are consistent with the theories postulating that overconfidence and mispricing due to biased expectations lead investors to hold financial positions that are not fully justified by the fundamentals of the companies and to react to news (that is, earnings announcements) by immediately correcting them (De Long et al., 1990, or Engelberg et al., 2018, among others). The result that the abnormal average potential values only slowly revert back to the average values historically observed when no earnings calls are scheduled is also consistent with the limited attention hypothesis: If investors are distracted and focus on irrelevant news instead of correctly processing the relevant new information included in the earnings releases, they underreact to the news, and it may take some time before the mispricing is fully corrected (Hirshleifer et al., 2009).

Second, we find that in most cases the sign of the potential returns corresponds to the one of the sentiment. This result again supports the limited attention hypothesis causing investors' underreaction to news that leads to prices that are too low after good news and too high after bad news. As a consequence, positive (negative) sentiment predicts high (low) subsequent returns (Hirshleifer et al., 2009). Third, we find a leverage effect in the estimated average potential outcomes: In case of negative sentiment, the potential return, volatility, and trading volume are larger (in magnitude) than those estimated for positive sentiment. This result conforms well with previous theoretical and empirical studies that found an asymmetric reaction in the size of the mispricing, volatility, and trading volume depending on the sign of the sentiment; see, among others, Miller (1977) and Zhang et al. (2016).

Fourth, the average estimated potential effects in response to mixed opinions and when there is disagreement in the news articles' sentiment are similar but smaller in magnitude than those found for negative sentiment; in fact, they lie between the values estimated for earnings announcements that cause only negative or only positive sentiments. This finding complements and partially contrasts with the empirical results shown in the previous literature that observed a negative (positive) relation between the divergence of investors' opinions and sentiments and the subsequent absolute returns (volatilities and trading volumes) (Siganos et al., 2017; Giannini et al., 2019; Cookson and Niessner, 2020). Similarly to what Antweiler and Frank (2004) and

Audrino et al. (2020) reported in a predictive setting, after controlling for other confounding factors in a causal framework, the impact of post-earnings announcement sentiment disagreement on subsequent returns, volatilities, and trading volumes is reduced such that the estimated potential outcomes become smaller in magnitude than those observed for negative sentiments shared across all news articles.

Fifth, we find that generally the average potential returns across the different sentiment types are not statistically significantly different from zero at the 1% significance level except when considering the difference between the average weekly potential returns in reaction to positive and negative sentiments: In this case the difference in the average effects amounts to 76 basis points and is also largely economically relevant. With respect to differences in the estimated average potential volatilities and trading volumes, the picture becomes clearer both in terms of statistical and economic significance: After positive sentiment the effects are significantly smaller than after negative and mixed sentiments for which no particular differences are found (up to 10% and 13% smaller at the daily and weekly frequency, respectively). Thus, earnings announcements that produce only positive sentiments are substantially more effective in reducing market risk when investors trade financial positions in the corresponding companies than when sentiments are negative or mixed.

Finally, we conduct a heterogeneity analysis for various underlying variables. We investigate whether the estimated effects and the difference among them change (i) if the general economy is in an expansion or recession cycle; (ii) for different levels of general market risk and aggregated investors' sentiment as measured by the VIX; and (iii) for different levels of media coverage attention around the earnings releases as measured by the number of news articles published related to the announcements. We find that in most cases the difference in the average effects for volatility and trading volume between positive and negative (or mixed) sentiments becomes more pronounced when the macroeonomic conditions are worse and investors' aggregate perception about the uncertainty in the financial market is larger. In contrast, we did not find any particular difference in the estimated average effects due to different types of sentiment with respect to a different degree of coverage attention to the earnings announcements: When the attention is high, all daily effects for returns, volatility, and trading volume are similarly larger (in magnitude) irrespectively of the sentiment's types, which supports the limited attention hypothesis (that is, the price adjustment speed is faster when attention is higher) and confirms the empirical results published in the previous literature; see, among others, Peng (2005), Hirshleifer et al. (2009), Andrei and Hasler (2015), or Audrino et al. (2020).

The remainder of the paper is structured follows. Section 2 describes the data sources and the design of the study in the causal inference framework. The methodology based on causal forests and the corresponding estimators are introduced in Section 3. Section 4 contains and discusses the results. Section 5 concludes the paper.

2 The data

To analyze the causal links between the firm-specific sentiment related to earnings announcements and financial outcomes, we consider 807 companies whose stock was part of the S&P 500 index at some point during the time period ranging between September 10, 2003 and December 31, 2017.

Data sources. A variety of data providers have been used to build our dataset which, as in any causal study, consists of treatment, outcome, and control variables. The news stories and the corresponding news sentiment proxies regarding earnings announcements used in our dataset were obtained from the RavenPack News Analytics (RPNA) database. Compustat was used to retrieve historical information at a daily frequency about the S&P 500 stocks under consideration. The IBES (Institutional Brokers' Estimate System) database was used to identify quarterly earnings announcements properly time stamped. Finally, the NYSE Trade and Quote (TAQ) database was used to compute realized volatilities, one of our outcome variables. Detailed information on how all these data sources interact in the construction of our dataset can be found in the Appendix A.1.

Treatment variables. For each earnings announcement of each S&P 500 company (within its duration in the index), we collect the earnings news within a time window ranging from the announcement time-stamp to the next market opening. We refer to this period as the treatment window. To measure the outcomes and controls, we define the pre- and post-treatment windows which include 8 calendar days (open to close), respectively. A sketch of the time line is depicted in Figure 1.

We note that corporate earnings are generally released on a quarterly basis. Moreover, in view of the distribution of the times of release during the day (see Figure 2), we only focus on the announcements outside the market trading hours.

Once the companies and the timing of their earnings announcements are identified, we turn to the creation of our treatment variable using the sentiment conveyed by related news reports. The treatment variable can take three different values: *positive*, *negative*, and *mixed*. This value is computed by aggregating the sentiment scores provided by RPNA associated with all the news stories that correspond to each earnings release and to each company. This aggregation is

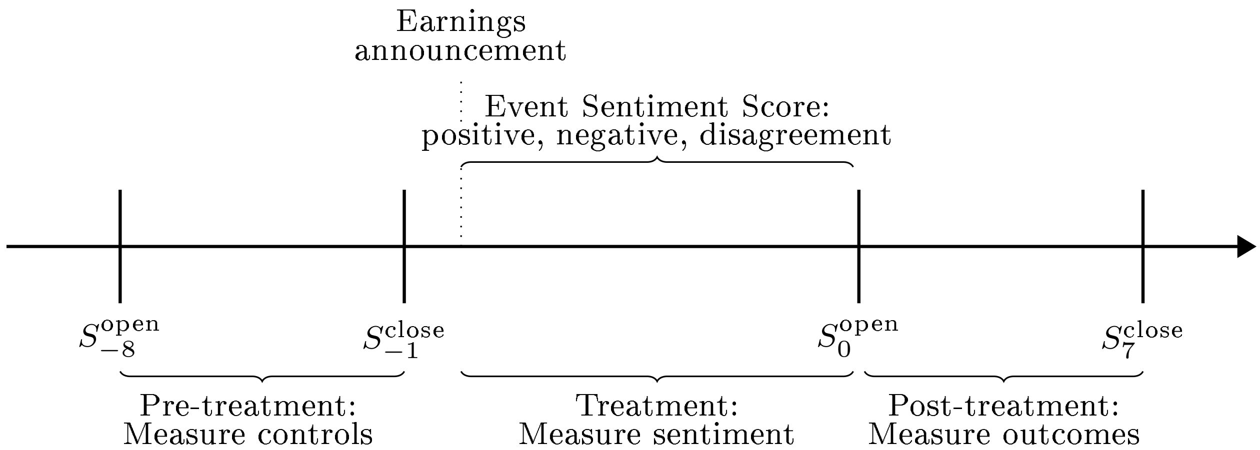

Figure 1: Sketch of the time line. For each earnings announcement, the treatment period (in which we collect the news stories about it) is defined as the time window ranging from the announcement time-stamp to the next market opening. The pre- and post-treatment windows include 8 calendar days (open to close) before and after it, respectively. The Event Sentiment Score (*ESS*) is a score ranging from 0 to 100 which represents the underlying sentiment for a news story and a given entity: A score of 50 indicates a neutral sentiment, while values above 50 indicate a positive sentiment and values below 50 a negative one; for more details see Appendix A.1.

carried out within the treatment window. When all news releases observed in this window have a non-negative sentiment, we assign the earnings release a *positive* treatment value. Similarly, if all news releases in the time frame exhibit a non-positive sentiment, the announcement is assigned a *negative* treatment value. Lastly, if we observe news stories with both negative and positive sentiments, we assign the release a *mixed* treatment value. We discard earnings releases for which we only observe neutral sentiments. More details about how we proxy the news sentiment using RavenPack can be found in Appendix A.1.

Table 1 presents the average, minimum and maximum numbers of news reports reported by RPNA per earnings announcement for the full sample and each sentiment category, respectively. On average, more news releases are observed if the earnings release is classified to be positive than negative, even though the mixed category is the most numerous.

Figure 3 shows the time-evolution of the sentiment variables. As is evident from the figure, news reports are classified by RPNA in the positive category much more often than in the other two classes. This is a consequence of the fact that company earnings reports are intentionally written in a way that try to convey as much as possible a positive feeling to the readers. Another prominent feature that is clearly visible in this graph is a structural break from 2009 to 2010, after which the share of negative sentiments abruptly decreased. We attribute this discontinuity to the sale in 2010 of the main RPNA information provider, namely the Dow Jones Indexes

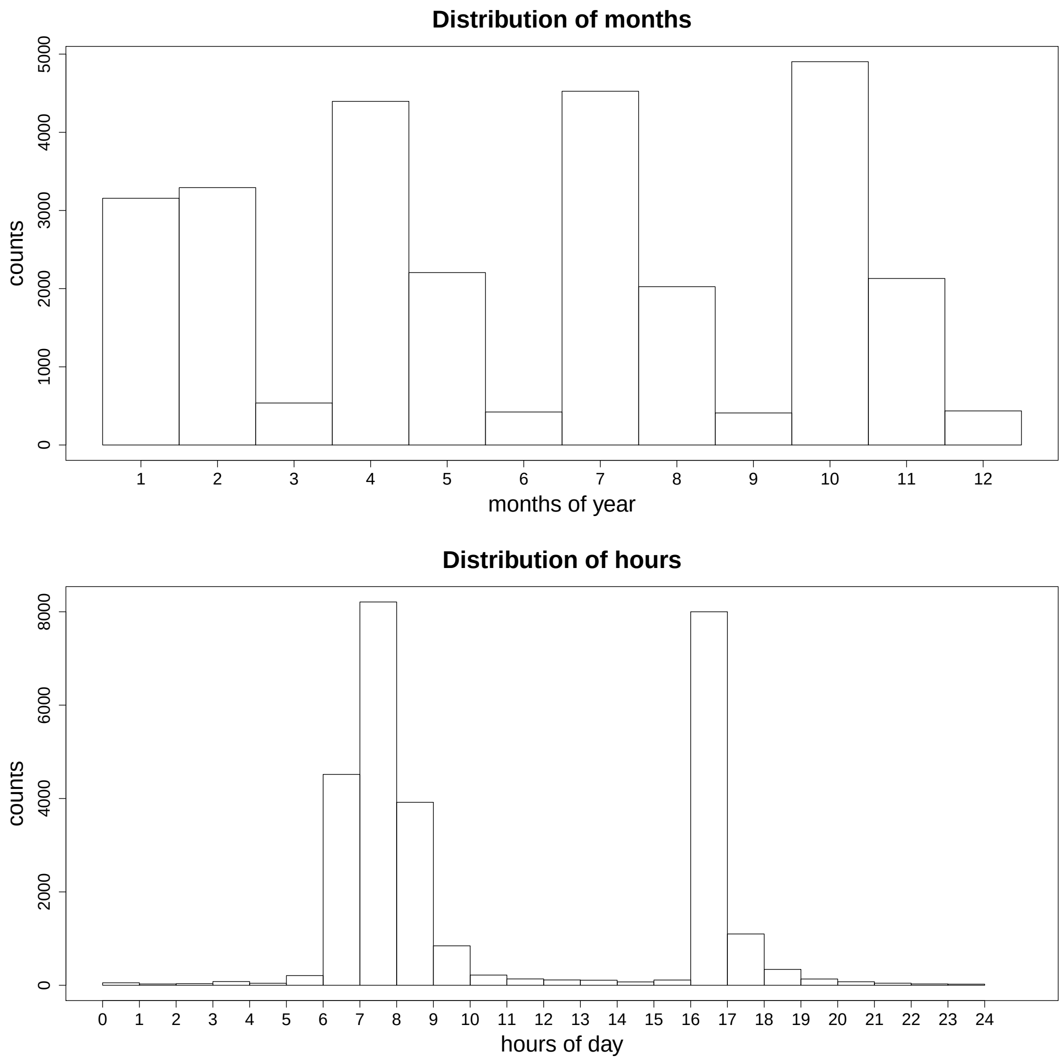

Figure 2: Distributions of earnings releases over the months of a year and the hours of a trading day. We consider all companies whose stock was listed in the S&P 500 index at some point between September 10, 2003 and December 31, 2017. Most announcements take place outside market trading hours.

subsidiary, to the CME Group. This operation carried in its wake a major shift in volume and focus in regard to their financial news publications. In order to account for these two different regimes later on in the estimation, a dummy variable that accounts for this structural break will be included in all analyses.

Outcome and control variables. Three types of financial outcome variables are simultaneously considered in our analysis: log-returns, realized volatilities, and the liquidity measured using the daily turnover ratios for each of the stocks. The three variables are measured during the post-treatment window. The volatility and liquidity measures are aggregated by averaging

	Full	Negative	Mixed	Positive
Mean	4.61	3.33	5.42	4.36
Min	1	1	2	1
Max	12	8	12	11

Table 1: News per earnings announcement. Average, minimum, and maximum numbers of news reports reported by RavenPack News Analytics (RPNA) per earnings announcement for all companies whose stock was listed in the S&P 500 index at some point between September 10, 2003 and December 31, 2017, and each sentiment category: All news stories observed share a non-negative sentiment (positive), all news stories exhibit a non-positive sentiment (negative), and if we observe news stories with both negative and positive sentiments (mixed).

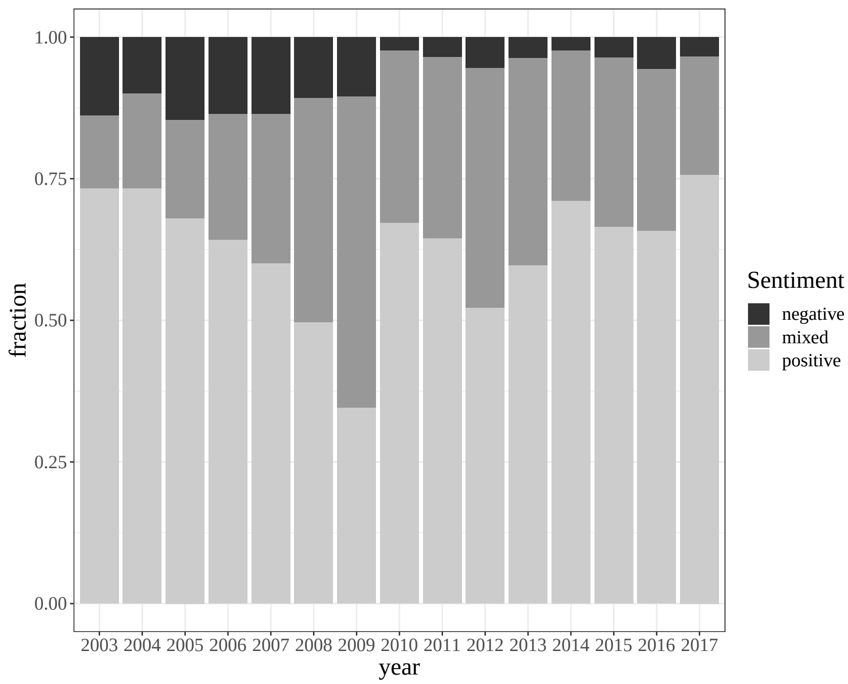

Figure 3: Distribution of sentiment categories over time. We consider all companies whose stock was listed in the S&P 500 index at some point between September 10, 2003 and December 31, 2017. We classify an earnings release to positive (negative) sentiment when all news stories observed share a non-negative (non-positive) sentiment, and to mixed sentiment if we observe news stories with both negative and positive sentiments. A structural break in 2009-2010 is clearly visible that coincides with the sale of the Dow Jones Indexes subsidiary to the CME Group.

the daily estimators over the one-week window. Additionally, the three outcomes measured over the one-day pre/post-treatment windows are also used as responses to examine the short-run effects. A detailed description of how we construct the outcome variables is deferred to Appendix A.2.

To control for the impact of other economic and financial variables, we include as covariates lagged outcomes, the global sector classification, the NBER-based recession indicator, the FOMC meeting indicator, the VIX, the Fama-French (FF) 3 factors, the CBOE Volatility Index (VIX), the US Gross Domestic Product (GDP), the Consumer Price Index (CPI) and some other relevant financial statement variables. Table A.1 summarizes the detailed definition and data sources for the financial and macroeconomic controls considered in our analysis. Table 2 shows the means and standard deviations of all outcome and control variables for the full sample and for the three sentiment categories separately. With respect to the response variables, a clear monotonic pattern is evident: Returns (realized volatilities or proxy for liquidity) are smaller (larger) when going from the positive to the negative sentiment categories, with values corresponding to mixed sentiments always lying in-between.

3 Empirical Strategy

3.1 The causal question and identification

As explained in the preceding sections, our objective is to determine the causal effects of post-earnings announcement sentiment on the dynamics of firms' stocks. In a causal inference framework, sentiment is encoded as a treatment variable that can take three different values: negative (n), mixed (m), and positive (p). Details on the multiple treatment version of the potential outcomes framework can be found in Imbens (2000) and Lechner (2001). Let $\mathcal{S} = \{n, m, p\}$ denote the three sentiment categories. In this setup, each firm f has for each announcement a three different potential outcomes $Y_{fa}(n)$, $Y_{fa}(m)$, and $Y_{fa}(p)$ (in our case these are daily returns, realized volatility, and a proxy for liquidity) that would be realized under the different sentiments. A causal effect is defined then as the contrast between two potential outcomes for different sentiments, $Y_{fa}(s) - Y_{fa}(s'),\ s, s' \in \mathcal{S}$. For example, the causal effect of mixed vs. negative sentiment is given by $Y_{fa}(m) - Y_{fa}(n)$. The most challenging part of causal inference is that only the potential outcome of the realized sentiment is observed. Indeed, if we write the observed outcome as $Y_{fa} = \mathbb{1}\{W_{fa} = n\}Y_{fa}(n) + \mathbb{1}\{W_{fa} = m\}Y_{fa}(m) + \mathbb{1}\{W_{fa} = p\}Y_{fa}(p)$, where $\mathbb{1}\{\cdot\}$ is the indicator function and W_{fa} denotes the treatment (sentiment) variable, it is clear that two out of the three potential outcomes are not accessible and hence the firm and announcement-specific effects are unobservable and not identified without imposing very restrictive assumptions. Instead, we follow the causal inference literature (e.g. Imbens and Rubin, 2015) and aim for aggregate estimands like the average treatment effect, $ATE_{s,s'} := E[Y_{fa}(s) - Y_{fa}(s')]$ or group average treatment effects, $GATE_{s,s'}(z) = E[Y_{fa}(s) - Y_{fa}(s') \mid Z = z]$, where the variable Z

	Negative		Mixed		Positive	
	Mean	SD	Mean	SD	Mean	SD
Outcomes:						
Log return (1 day) * 100	-0.17	4.76	-0.09	3.82	-0.05	3.01
Log return (1 week) * 100	-0.66	8.67	-0.22	6.74	0.10	5.01
Volatility (1 day) * 100	2.27	1.68	1.90	1.49	1.59	0.90
Volatility (1 week) * 100	2.40	1.69	2.01	1.29	1.67	0.81
Turnover ratio (1 day) * 100	3.50	4.10	2.90	3.23	2.40	2.72
Turnover ratio (1 week) * 100	2.07	2.24	1.66	1.51	1.36	1.18
Controls:						
Log return (1 day) lag * 100	0.11	2.69	0.03	2.19	0.08	1.69
Log return (1 week) lag * 100	0.16	7.52	0.26	5.41	0.43	4.03
Volatility (1 day) lag * 100	0.02	0.02	0.02	0.01	0.02	0.01
Volatility (1 week) lag * 100	2.09	1.55	1.75	1.27	1.44	0.81
Turnover ratio (1 day) lag * 100	1.77	1.93	1.50	1.81	1.25	1.19
Turnover ratio (1 week) lag * 100	1.52	1.69	1.21	1.11	1.00	0.82
NBER recession	0.16	0.37	0.16	0.36	0.07	0.26
FOMC meeting indicator	0.22	0.41	0.25	0.43	0.25	0.43
Loading on the market risk	1.16	0.58	1.04	0.46	1.03	0.52
Log of total assets	9.55	1.36	9.84	1.40	9.61	1.34
Market to book ratio	2.89	15.33	2.84	36.85	4.47	110.46
VIX lag	18.56	9.08	20.30	10.66	17.49	7.94
GDP	4.03	3.35	3.22	3.23	4.24	2.55
CPI	2.37	1.54	1.91	1.57	2.11	1.29
Observations	1837		7634		15666	

Table 2: Means and standard deviations of all outcome and control variables. We consider all companies whose stock was listed in the S&P 500 index at some point between September 10, 2003 and December 31, 2017. We classify an earnings release to positive (negative) sentiment when all news stories observed share a non-negative (non-positive) sentiment, and to mixed sentiment if we observe news stories with both negative and positive sentiments. The denomination "Turnover ratio" corresponds to the proxy for the liquidity constructed using the daily turnover ratios for the relevant stocks. Notice that lagged versions of the outcome variables are included as controls.

defines a subgroup of interest to compare, for example, the effects in and out of recessions or for different industrial sectors.

The identification of the ATE and the GATE requires the observation of control variables X that are simultaneously related to the sentiment about an earnings announcement and to the outcome of interest. Given this set of variables, the most detailed effect that we could consider is the individualized average treatment effects, $IATE_{s,s'}(x) = E[Y_{fa}(s) - Y_{fa}(s') \mid X = x]$. We consider in our study the control variables shown in the second part of Table 2. The macroeconomic and financial variables contained in that list have been empirically shown to contain information about the time-varying first and second conditional moments of the return dynamics described by our outcome variables; see, among others, Christiansen et al. (2012), Mittnik et al. (2015), Paye (2012), Nonejad (2017), and references therein for justifications of this statement. Additionally, we control for unobserved temporal and sector-specific effects by including the reporting season and sectors in the analysis.

There are other covariates that regularly appear in the literature in connection with the forecasting of our outcome variables and that belong to three main families, namely, interest rates, macroeconomic variables (e.g. inflation, unemployment, monetary masses, consumer and producer sentiment surveys), and commercial and industrial activity indicators (e.g. oil and energy prices, Baltic Index, Industrial Production Index). Since these three families are directly or indirectly represented in our list of confounders, we are confident that we capture the most important controls. Nevertheless, some of this missing information could still bias our results.

3.2 Estimator

In this paper, we utilize the results made available in the recent causal machine learning literature. (For an only slightly outdated survey, see Athey and Imbens, 2017; see also Athey, 2017 for a more general motivation of this research subject.) This approach combines the prediction power of machine and statistical learning (for an overview see, e.g., Hastie et al., 2009) with results in the microeconometric literature for the definition and identification of causal effects (e.g., Rubin, 1974, and Imbens and Wooldridge, 2009, for a recent overview). This field has seen a recent surge of proposed methods, in particular in epidemiology and econometrics. Knaus et al. (2020) compare systematically many of those methods with respect to their theoretical properties as well as their performance in a simulation exercise. One conclusion from this paper is that random forest-based estimation approaches seem to outperform many alternative estimators.

The starting point of the causal forest literature is the causal tree introduced in a paper by Athey and Imbens, 2016. In a causal tree, the sample is split sequentially into smaller and smaller strata, in which the values of the covariates X become more and more homogeneous, in order to mitigate selection effects and to uncover effect heterogeneity. Once the splitting is terminated based on some stopping criterion, the treatment effect is computed within each stratum (called a "leaf") by computing the difference of the mean outcomes of treated and controls (possibly weighted by the conditional on X probabilities of being a treated or control observation). However, the literature on regression trees acknowledges that the sample splits may be rather unstable because of their sequential nature (if the first split is different, the full tree will likely lead to different final strata). A solution to this problem is so-called random forests. Their key idea is to induce some randomness into the tree building process, build many trees, and then average the predictions of the many trees. The induced randomness is generated by using randomly generated subsamples (or bootstrap samples) and by considering for each splitting decision only a random selection of the covariates. Wager and Athey (2018) use this idea to propose causal forests, which are based on a collection of causal trees with small final leaves.[1] Lechner (2018) develops these ideas further by improving on the splitting rule for the individual trees, and by providing methods to estimate heterogeneous effects for a limited number of discrete policy variables (Group Average Treatment Effects, GATE) at low computational costs, in addition to the highly disaggregated effects the literature has focused on thus far (Individualized Average Treatment Effects, IATE).

Furthermore, his paper suggested a way of performing unified inference for all aggregation levels. Finally, the approach is applicable to a multiple, discrete treatment framework. Since many of these advantages are important in the empirical analysis of this paper, this approach is used below.

3.3 Details of the Modified Causal Forest (MCF) estimator

Denoting the conditional expectations of Y given X for the sentiment category $W = s$ by $\mu_s(x)$, i.e. $\mu_s(x) = E[Y(s) \mid X = x, W = s]$, leads to the following expression of the $\text{IATE}_{s,s'}(x)$ as a difference of $\mu_s(x)$ and $\mu_{s'}(x)$:[2]

$$\text{IATE}_{s,s'}(x) = \mu_s(x) - \mu_{s'}(x) = E[Y(s) - Y(s') \mid X = x], \ \forall \ x \in \mathcal{X}, \ \forall \ s \neq s' \in \mathcal{S}.$$

This estimation task is different from standard machine learning (ML) problems because the two conditional expectations have to be estimated in different, treatment-specific subsamples.[3]

An easy-to-implement estimator consists of estimating the two conditional expectations separately by standard ML tools, and taking a difference. However, this approach has the disadvantage that standard ML methods attempt to maximize out-of-sample predictive power of the two estimators $s$eparately. More concretely, if Random Forest is used, the difference between the

[1] Athey et al. (2019) generalize this idea to many different econometric estimation problems.

[2] For the ease of exposition we do not write explicitly the subscript fa in this section.

[3] This is implied by the fact that the causal effect is a hypothetical construct that is per se unobservable. Thus, in the words of Athey and Imbens (2016), the "ground truth" is unobservable in causal analysis.

predictions of the two different estimated forests (one estimated in subpopulation s, the other one estimated in sub-population s') may suggest a variability of the IATE that is just estimation error due to (random) differences in the estimated forests. This problem can be particularly pronounced when the features are highly predictive for Y, but not for the IATEs.

An alternative approach is to use the same trees in both subsamples in which $\mu_s(x)$ and $\mu_{s'}(x)$ are estimated by $\widehat{\mu}_s(x)$ and $\widehat{\mu}_{s'}(x)$, respectively. Of course, the key is then how to obtain a plausible splitting rule for this "joint" forest (that leads to a correlation of $\widehat{\mu}_s(x)$ and $\widehat{\mu}_{s'}(x)$). Lechner (2018) suggests a splitting rule that considers the mean square error of this estimation problem directly. For the comparison between two alternatives at a given value of x, we obtain:

$$\begin{aligned}\mathrm{MSE}[\widehat{\mathrm{IATE}}_{s,s'}(x)] &= E\Big[\big(\widehat{\mathrm{IATE}}_{s,s'}(x) - \mathrm{IATE}_{s,s'}(x)\big)^2\Big] \\ &= E\big[(\widehat{\mu}_s(x) - \mu_s(x))^2\big] + E\big[(\widehat{\mu}_{s'}(x) - \mu_{s'}(x))^2\big] - 2E\big[(\widehat{\mu}_s(x) - \mu_s(x))(\widehat{\mu}_{s'}(x) - \mu_{s'}(x))\big] \\ &= \mathrm{MSE}[\widehat{\mu}_s(x)] + \mathrm{MSE}[\widehat{\mu}_{s'}(x)] - 2\mathrm{MCE}[\widehat{\mu}_s(x), \widehat{\mu}_{s'}(x)],\end{aligned}$$

with $\mathrm{MCE}[\widehat{\mu}_s(x), \widehat{\mu}_{s'}(x)] = E\big[(\widehat{\mu}_s(x) - \mu_s(x))(\widehat{\mu}_{s'}(x) - \mu_{s'}(x))\big]$. When there are more than two treatments (as in our application), this is averaged over the treatments and over the observations used for forest building:

$$\mathrm{MSE}(x) = \sum_{s \neq s'} \mathrm{MSE}[\widehat{\mathrm{IATE}}_{s,s'}(x)] = (\#\mathcal{S} - 1) \sum_{s \in \mathcal{S}} \mathrm{MSE}[\widehat{\mu}_s(x)] - 2 \sum_{s \neq s'} \mathrm{MCE}[\widehat{\mu}_s(x), \widehat{\mu}_{s'}(x)],$$

where $\#\mathcal{S}$ denotes the number of possible treatments in $\mathcal{S}$.

For constructing estimators based on this criterion, the MSE of $\widehat{\mu}_s(x)$ has to be estimated. This is straightforward, as the MSEs of all functions $\widehat{\mu}_s(x)$, $s \in \mathcal{S}$, can be computed in the respective treatment subsamples in the usual way. Denote by N and $N^s_{S_x}$ the total number of observations and the number of observations with treatment value s in a certain stratum (leaf) S_x, which is defined by the values of the features x, respectively. Then, the following estimator is a "natural" choice:

$$\widehat{\mathrm{MSE}}_{S_x}[\widehat{\mu}_s(x)] = \frac{1}{N^s_{S_x}} \sum_{i=1}^{N} \mathbb{1}\{x_i \in S_x\} \mathbb{1}\{w_i = s\} (\widehat{\mu}_s(x_i) - y_i)^2$$

Note that the overall MSE in S_x is the sum of the MSEs in the treatment-specific subsamples of S_x, where each subsample receives the same weight (independent of the number of observations in that subsample), as implied by the above MSE formula for causal effect estimation.

In order to compute the correlation of the estimation errors, we need a proxy for cases when there are no observations with exactly the same values of x in all treatment states (as is always true for continuous features). In this case, we propose using the closest neighbor available (which is denoted by $y(i,s)$ below) instead:[4]

$$\widehat{\text{MCE}}[\widehat{\mu}_s(x), \widehat{\mu}_{s'}(x)] \\ = \frac{1}{N^s_{S_x} + N^{s'}_{S_x}} \sum_{i=1}^{N} \mathbb{1}\{x_i \in S_x\}(\mathbb{1}\{w_i = s\} + \mathbb{1}\{w_i = s'\})(\widehat{\mu}_s(x_i) - \tilde{y}_{i,s})(\widehat{\mu}_{s'}(x_i) - \tilde{y}_{i,s'}),$$

where $\tilde{y}_{i,s} = y_i$ if $w_i = s$ and the closest neighbor $y_{i,s}$ otherwise.

The splitting rule that minimizes $\widehat{\text{MSE}}[\widehat{\text{IATE}}_{s,s'}(x)]$ is motivated by maximizing the predictive power of the estimator. However, in causal analysis, inference is important. Thus, if the MSE-minimal estimator has a substantial bias, this is problematic. In causal studies, a substantial source of bias is a non-random allocation of treatment (selection bias). Lechner (2018) proposes to add a penalty term to $\widehat{\text{MSE}}[\widehat{\text{IATE}}_{s,s'}(x)]$ that penalizes possible splits where the treatment probabilities in the resulting daughter leaves are similar; in fact, splits leading to leaves with similar treatment shares will not be able to remove much selection bias, whereas if they are very different in this respect, they approximate differences in $P[W = s \mid X = x]$ well. In other words, the modified criterion prefers splits with high propensity score heterogeneity and puts explicit emphasis on tackling selection bias. In our study we proceed along these lines and add to the splitting criteria the following penalty function suggested in Lechner (2018):

$$I(\lambda, x^{'}, x^{''}) = \lambda\Big[1 - \frac{1}{\#\mathcal{S}} \sum_{s \in \mathcal{S}} \big[P[W = s \mid X \in \text{leaf}(x^{'})] - P[W = s \mid X \in \text{leaf}(x^{''})]\big]^2\Big],$$

where $\text{leaf}(x^{'})$ and $\text{leaf}(x^{''})$ denote the values of the features in the daughter leaves resulting from splitting some parent leaf and λ is a tuning parameter which is set to equal the variance of Y.

Estimates for GATEs and ATEs are most easily obtained by averaging the IATEs in the respective subsamples defined by z (assuming discrete Z) and $\Delta = \{s, s'\}$. Thus, we estimate the GATEs and ATEs as appropriate averages of the $\widehat{\text{IATE}}_{s,s'}(x)$:

$$\begin{aligned} \widehat{\text{GATE}}_{s,s'}(z) &= \frac{1}{N^{z,\Delta}} \sum_{i=1}^{N} \mathbb{1}\{z_i = z, w_i \in \Delta\}\widehat{\text{IATE}}_{s,s'}(x_i) \text{ and} \\ \widehat{\text{ATE}}_{s,s'} &= \frac{1}{N^{\Delta}} \sum_{i=1}^{N} \mathbb{1}\{w_i \in \Delta\}\widehat{\text{IATE}}_{s,s'}(x_i), \end{aligned}$$

where $N^{z,\Delta} = \sum_{i=1}^{N} \mathbb{1}\{z_i = z, w_i \in \Delta\}$ and $N^{\Delta} = \sum_{i=1}^{N} \mathbb{1}\{w_i \in \Delta\}$.

[4] Closeness is based on a simplified Mahalanobis metric as in Athey and Imbens (2016). This simplified version has the inverse of the variances of the features on the main diagonal. Off-diagonal elements are zero. The simplification avoids computational complications when inverting the variance-covariance matrix of potentially large-dimensional features at the cost of ignoring correlations between covariates.

Under certain regularity conditions the resulting estimators for the causal effects are (pointwise) consistent and asymptotically normal. Exploiting the fact that Random Forest estimators are linear operators and can be represented as weighted averages of the outcomes, we can conveniently formulate one inference principle for all causal parameters of interest (always based on the same forest). In fact, it turns out that the variance of the estimator of interest (which could be one of the IATEs, or GATEs, or ATEs) can be computed in closed form and can be estimated with accurate results by standard non-parametric or ML methods. For any further details about the estimation and inference procedures, we refer to Lechner (2018).

3.4 Implementation

The main elements of the algorithm used for estimating the results are the following:

1) Split the estimation sample randomly into two parts of equal size (sample A and sample B).

2) Estimate the trees that define the respective causal forest in sample A. Estimate the same forest for all treatment states jointly. Before building the trees, for each observation in each treatment state, find a close neighbor (the same observation might be used as a neighbor several times) in every other treatment state and save its outcome (to estimate MCE). The splitting rule is based on minimizing the overall MSEs, taking into account the sum of all MCEs and an appropriate penalty function.

3) Apply the sample splits obtained in sample A to all subsamples (by treatment state) of sample B and take the mean of the outcomes in the respective leaf as the prediction that comes with this forest.

4) Obtain the weights from the estimated forest by counting how many times an observation in sample B is used to predict IATE for a particular value of x.

5) Aggregate the IATEs to GATEs by taking the average over observations in sample B that have the same value of z and treatment group Δ. Moreover, average over all observations in the treatment group Δ to obtain the ATEs.

6) Exploit the properties of linear operators to compute standard errors. Use the estimated standard errors together with the quantiles from the normal distribution to obtain critical values (p-values).

4 Results

In the next two sections we present the results obtained using the methodology introduced in Section 3 and using the data and the empirical design spelled out in Section 2.

4.1 Average effects

The estimated average effects on the three outcome variables chosen in Section 2, that is, returns, volatilities, and liquidity (proxied by the daily turnover ratio) are provided in Tables 3, 4, and 5, respectively. A common feature of the results regarding these three variables and for any value of the treatment is that biased expectations drive large nominal levels in the day immediately after the announcements that revert back to average values in the week that follows them. This finding is particularly noticeable for returns in which the jump may be of up to 75 basis points the trading day after the earnings announcement. Levels that are more than 60% larger than usual are also observed for volatility and liquidity the trading day after the announcement. This is the result of the changes investors have to make to their portfolios to correct wrong positions due to overconfidence and mispricing after the release of the earnings information. Moreover, abnormal values revert to the average values only after some days, indicating that a non-negligible part of the investors are distracted from other news releases and underreact to the new earnings information or react with some delay. This empirical result is consistent with the widely postulated limited attention hypothesis.

This feature is asymmetric for returns with respect to the sentiment class, as it is for negative sentiment where this short-run jump is the most visible. On the other hand, the short-run effect for volatility and turnover ratios is relatively homogeneous with respect to the different sentiment values. More generally, a leverage effect in the average potential outcomes is clearly visible as returns, volatilities, and turnover ratios are larger in case of negative sentiment than those estimated for positive sentiment. This observation is in agreement with well-documented stylized facts of the dynamics of financial time series.

Regarding the average treatment effects, it is only the difference between the positive and the negative sentiments and that of the positive and the mixed sentiments that are in most cases statistically significant at the 1% level for the three outcome variables. In those two cases, the effects observed go in the same direction but are larger in magnitude for negative sentiment. In partial contrast with previous empirical studies, we do not find that the positive (negative) relation between the presence of mixed investors' opinions and sentiments after an earnings release and the risk (return) of the stock is the largest. In fact, shared negative sentiments have a bigger impact on the stock price dynamics than mixed or positive sentiments. This new result is a consequence of the causal setting in which we perform our analysis where we properly control for other confounders.

The figures obtained are of particular economic significance as they quantify the potential for excess returns and (market and liquidity) risk mitigation associated with positive sentiment as opposed to a negative or a mixed one. For instance, when it comes to returns, there is a difference of an average of 76 basis points over a week in the returns recorded after an earnings announcement perceived as positive compared to a negative one. A similar phenomenon occurs from the point of view of market risk, as realized volatility reduces by 16 points on average over the successive week when comparing a positive with a negative treatment.

	Log return (1 day) * 100			Log return (1 week) * 100		
	Estimate	Std.	p-val in %	Estimate	Std.	p-val in %
Level:						
Negative	-0.214	0.138		-0.744	0.234	
Mixed	-0.053	0.055		-0.159	0.096	
Positive	-0.105	0.038		0.018	0.064	
Effects:						
Mixed - negative	0.161	0.148	27.7	0.585	0.253	2.1
Positive - mixed	-0.051	0.067	44.1	0.176	0.116	12.7
Positive - negative	0.109	0.143	44.3	0.762	0.243	0.2

Table 3: Average potential levels and treatment effects (ATEs) for daily and weekly (log-)returns. We consider all companies whose stock was listed in the S&P 500 index at some point between September 10, 2003 and December 31, 2017. We classify an earnings release to positive (negative) sentiment when all news stories observed share a non-negative (non-positive) sentiment, and to mixed sentiment if we observe news stories with both negative and positive sentiments.

4.2 Heterogeneous effects

The results in Table 6 show whether the average treatment effects differ during recession periods or as a function of the media coverage attention that a specific earnings announcement receives measured by the number of news articles published related to the announcements. Regarding the last point, we define the attention received by an earnings announcement as high when at least five news stories are observed in relation to it.[5]

[5] The cut-off of five corresponds to the median number of news stories.

	Volatility (1 day) * 100			Volatility (1 week) * 100		
	Estimate	Std.	p-val in %	Estimate	Std.	p-val in %
Lagged average levels:	1.73			1.58		
Levels:						
Negative	2.93	0.07		1.94	0.04	
Mixed	2.85	0.03		1.85	0.02	
Positive	2.72	0.02		1.78	0.02	
Effects:						
Mixed - negative	-0.075	0.079	34.5	-0.091	0.045	4.5
Positive - mixed	-0.132	0.035	0.0	-0.071	0.022	0.1
Positive - negative	-0.207	0.078	0.8	-0.161	0.045	0.0

Table 4: Average potential levels and treatment effects (ATEs) for daily and averaged weekly realized volatilities.

	Turnover ratio (1 day) * 100			Turnover ratio (1 week) * 100		
	Estimate	Std.	p-val in %	Estimate	Std.	p-val in %
Lagged average levels:	1.37			1.10		
Levels:						
Negative	2.85	0.10		1.66	0.05	
Mixed	2.72	0.05		1.53	0.02	
Positive	2.55	0.04		1.45	0.02	
Effects:						
Mixed - negative	-0.125	0.111	26.4	-0.127	0.050	1.1
Positive - mixed	-0.173	0.060	0.4	-0.082	0.027	0.3
Positive - negative	-0.297	0.107	0.5	-0.209	0.048	0.0

Table 5: Average potential levels and treatment effects (ATEs) for daily and averaged weekly turnover ratios as proxies for liquidity.

The results summarized in the table show that the effects differ substantially during different business cycles: Effects are generally much larger during recessions than expansion periods. This difference is particularly evident and statistically significant for realized volatilites. For instance, receiving only positive instead of only negative news reduces daily volatility during recessions by 0.93, which is nearly eight times more than the figure obtained in periods with positive economic growth (-0.12).

	Recession			Attention		
	no	yes	Diff.	low	high	Diff.
*Log return (1 day) * 100*:						
Mixed - negative	0.11	0.57	-0.46	0.17	0.15	0.01
	(0.13)	(0.55)	(0.54)	(0.14)	(0.18)	(0.16)
Positive - mixed	-0.03	-0.23	0.20	-0.02	-0.08	0.06
	(0.06)	(0.23)	(0.20)	(0.07)	(0.07)	(0.06)
Positive - negative	0.08	0.34	-0.26	0.15	0.07	0.07
	(0.13)	(0.54)	(0.53)	(0.13)	(0.18)	(0.15)
*Log return (1 week) * 100*:						
Mixed - negative	0.60***	0.49	0.11	0.73***	0.45	0.27
	(0.22)	(1.10)	(0.98)	(0.24)	(0.33)	(0.27)
Positive - mixed	0.16	0.29	-0.13	0.14	0.21*	-0.08
	(0.11)	(0.38)	(0.35)	(0.13)	(0.13)	(0.11)
Positive - negative	0.76***	0.78	-0.02	0.87***	0.67**	0.20
	(0.20)	(1.09)	(0.96)	(0.22)	(0.32)	(0.26)
*Volatility (1 day) * 100*:						
Mixed - negative	-0.03	-0.45	0.42	-0.08	-0.07	-0.01
	(0.08)	(0.33)	(0.37)	(0.09)	(0.16)	(0.21)
Positive - mixed	-0.09***	-0.48**	0.39**	-0.11*	-0.15***	0.04
	(0.03)	(0.20)	(0.20)	(0.07)	(0.05)	(0.09)
Positive - negative	-0.12	-0.93***	0.81**	-0.19**	-0.22	0.03
	(0.08)	(0.33)	(0.36)	(0.08)	(0.16)	(0.21)
*Volatility (1 week) * 100*:						
Mixed - negative	-0.06	-0.32	0.26	-0.09*	-0.09	0.01
	(0.04)	(0.23)	(0.25)	(0.05)	(0.10)	(0.14)
Positive - mixed	-0.04***	-0.30**	0.26*	-0.06	-0.08***	0.03
	(0.02)	(0.14)	(0.13)	(0.04)	(0.03)	(0.06)
Positive - negative	-0.11**	-0.62***	0.52**	-0.14***	-0.18*	0.03
	(0.04)	(0.23)	(0.24)	(0.04)	(0.10)	(0.13)
*Turnover ratio (1 day) * 100*:						
Mixed - negative	-0.10	-0.32	0.22	-0.07	-0.17	0.10
	(0.11)	(0.25)	(0.28)	(0.12)	(0.18)	(0.20)
Positive - mixed	-0.17***	-0.19	0.02	-0.19**	-0.16**	-0.04
	(0.06)	(0.17)	(0.15)	(0.08)	(0.07)	(0.08)
Positive - negative	-0.27***	-0.51**	0.24	-0.26**	-0.33*	0.06
	(0.11)	(0.25)	(0.27)	(0.10)	(0.17)	(0.19)
*Turnover ratio (1 week) * 100*:						
Mixed - negative	-0.10**	-0.34**	0.24	-0.11**	-0.14*	0.03
	(0.05)	(0.15)	(0.17)	(0.05)	(0.09)	(0.11)
Positive - mixed	-0.08***	-0.13	0.05	-0.08**	-0.09***	0.01
	(0.02)	(0.09)	(0.08)	(0.04)	(0.03)	(0.04)
Positive - negative	-0.18***	-0.47***	0.29*	-0.19***	-0.23***	0.04
	(0.05)	(0.15)	(0.17)	(0.05)	(0.09)	(0.10)

Table 6: Heterogeneity analysis for different business cycles and media coverage attention values. Attention of a specific earnings announcement is measured by the number of news articles published related to it: We define the attention received by an earnings announcement as high when at least five news stories (median value) are observed in relation to it.

These results show that average treatment effects across different investors' sentiments are exacerbated during bad economic cycles, periods in which investors' behavioral biases tend to be more pronounced. This interpretation is also in line with the results that we get when performing a heterogeneity analysis with respect to different VIX values as shown in Figures 4 and 5. These figures compare groups' average treatment effects (GATEs) on volatilities and turnover ratios obtained for 21 equally sized bins of VIX to the average treatment effect (ATE) and document that the effects are amplified if uncertainty in the general financial market grows.[6] In fact, for the groups corresponding to higher values of the VIX, differences between the ATE and the GATE become larger (in magnitude) and significantly more negative.

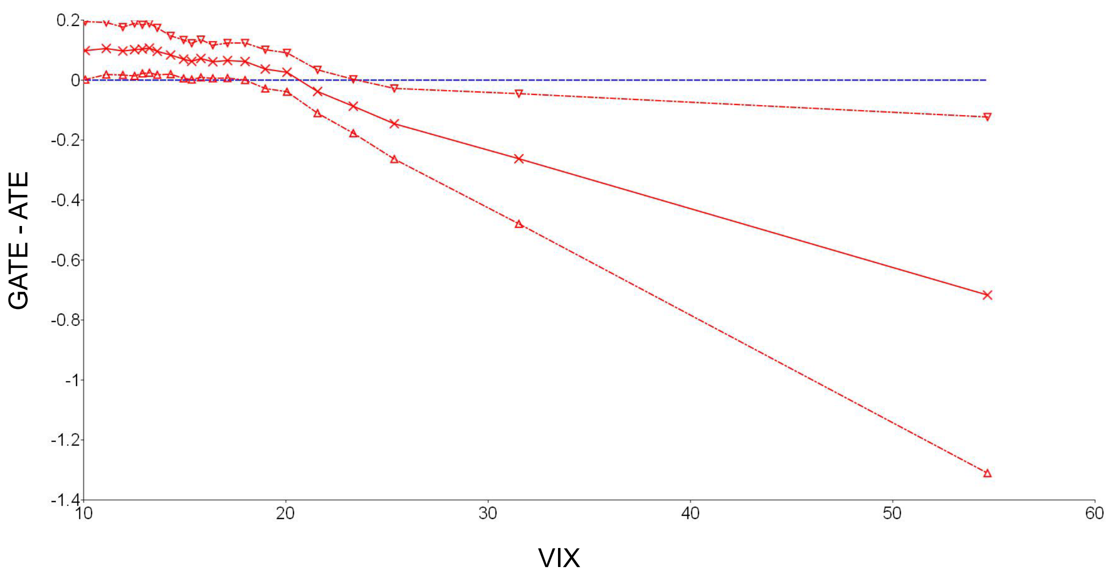

Figure 4: Difference between the group average treatment effects (GATEs) and the average treatment effect (ATE) on realized volatilities of positive vs. negative sentiment. Groups are constructed splitting the observed VIX values into 21 (equally sized) classes. The respective 90% confidence intervals are also drawn.

In contrast, results summarized in Table 6 show that different levels of attention have no statistically significant influence on the size of the treatment effects. In fact, when coverage attention is high, all daily effects become similarly larger in magnitude than when attention is low irrespectively of the sentiment's category. When the media coverage of an earnings announcement is high, investors seem to react in a similar, more timely manner to all types of information (positive or negative), adjusting their positions.

[6]The respective graphs of the other 16 effect-outcome combinations can be obtained from the authors upon request.

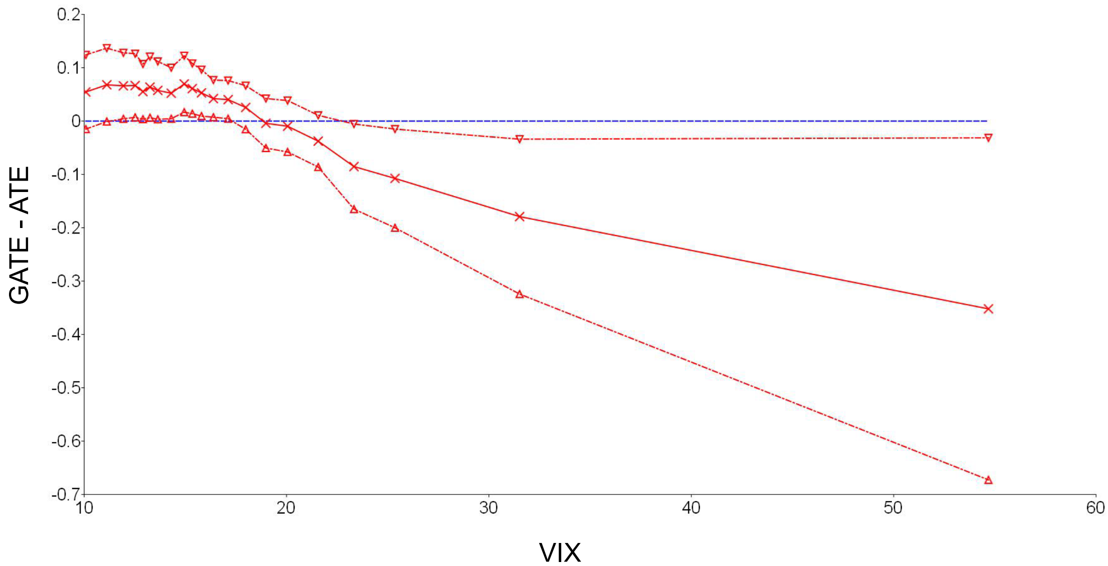

Figure 5: Difference between the group average treatment effects (GATEs) and the average treatment effect (ATE) on turnover ratios of positive vs. negative sentiment. Groups are constructed splitting the observed VIX values into 21 (equally sized) classes. The respective 90% confidence intervals are also drawn.

5 Conclusion

We studied the impact on the dynamics of stock prices of post-earnings announcement sentiment in relation to corporate earnings announcements on the dynamics of stock prices. Even though this problem has already been treated in the literature due to its obvious strategic importance, our work incorporates two main novelties. First, our approach is based on novel causal machine learning-based techniques, namely causal forests enhanced with a modified estimator of the treatment effect. Second, our results have been obtained using a large firm-specific dataset in which the financial outcomes have been computed out of high-frequency intra-day traded prices and the sentiments have been compiled and extracted out of all the news stories in relation to all S&P 500 firms' earnings announcements for fifteen years released by all the major financial news providers.

The results obtained with this new methodology and rich dataset confirm and quantify previous empirical observations in relation to this problem. In order to do so, we have first characterized the post-earnings announcement sentiment in our causal inference framework as a treatment variable that can take three values (positive, negative, and mixed sentiments) and we have chosen as outcome variables the impact on returns, volatilities, and liquidity (proxied by turnover ratios). The estimated average effects show that there is a noticeable short-run correction effect on returns mostly in the direction of the sentiment in the day that immediately follows the announcement. Returns go back to average values in the week that follows the announcement and correct the misspricing committed due to biased expectations. This effect is particularly visible for negative sentiments.

Regarding the average treatment effects, it is only the difference between the positive and negative sentiments and the difference between the positive and mixed sentiments that are mostly statistically significant at the 1% level for the three outcome variables. The figures that we obtained are of large economic significance as they quantify the potential for excess returns and (market and liquidity) risk mitigation associated with positive sentiment as opposed to a negative or a mixed one. For returns, there is a difference of an average of 76 basis points in the returns recorded after an earnings announcement perceived as positive as opposed to a negative one. A similar phenomenon occurs from the point of view of market risk, as realized volatility declines by 16 points when comparing a positive with a negative treatment.

Finally, a heterogeneity study was conducted to determine if recessions, general market uncertainty, and attention (measured by the number of news articles published in relation to the announcements) have an influence on the results. We have seen that this is not the case for attention, while we showed that recessions and high market uncertainty significantly exacerbate the effects.

Data Appendix

A.1 Ravenpack and news sentiment measures

The news stories and the corresponding news sentiment proxies regarding earnings announcements used in our dataset have been obtained from the *RavenPack News Analytics 4.0 - Dow Jones equities edition* database. RavenPack News Analytics (RPNA) automatically gathers and analyzes news releases from *Dow Jones Financial Wires*, the *Wall Street Journal*, *Barron's*, and *MarketWatch*, offering a database with a multitude of data fields for each news item captured.[7] A record in the RPNA database refers to a combination of a news story and a so-called entity.[8] When multiple entities are mentioned within a specific news story, a record is created for each entity mentioned in it.

We filter and prepare our dataset by focusing on the event group (GROUP), the event novelty score (ENS), the relevance (RELEVANCE), the CUSIP code (CUSIP), and the event sentiment score (ESS). We briefly describe these concepts in the following paragraphs.

- The GROUP is the second highest level of the RavenPack event taxonomy. It refers to a collection of related events and is automatically tagged from the headline and news content. As we only focus on news related to earnings announcements, we filter the data using this field, and hence all our data belong to the *earnings* GROUP.

- The ENS is a score ranging from 0 to 100, representing how novel a news story is within a 24-hour time window. Two stories about the same event for the same entity within a 24-hour period are considered to belong to the same chain of events. Chains of events have a decaying ENS; the first news story relating to a specific event and entity within a 24-hour time frame will be assigned a score of 100, while any subsequent story in the same chain of event will have a lower score[9]. If no news is published within 24 hours, the ENS is reset. We focus on the most novel news and filter our data to keep only stories with an ENS of 100.

- The RELEVANCE is a score between 0 and 100, indicating how strongly an entity is related to the underlying news story. A high RELEVANCE score implies that the identified entity plays a major role in the story, while lower scores refer to more passive mentions of the entities.

[7] A detailed overview of the fields in the RavenPack database can be found in Mitra and Mitra (2011).

[8] As we focus on the equities edition, all entities are companies in our case.

[9] The ENS follows a decay function with values 100, 75, 56, 42, 32, 24, 18, 13, 10, 7, 6, 4, 3, 2, 2, 1, 1, 1, 1, 0, ...

- The CUSIP (Committee on Uniform Securities Identification Procedures) is a well-known coding system for US stocks. We use this identifier to link data coming from RPNA and the Compustat and IBES databases. Compustat is used in our study to retrieve at a daily frequency historical information about the stocks that make up the S&P 500 index. The IBES database provides the corresponding quarterly earnings announcement properly time stamped. More specifically, for each earnings announcement of each S&P 500 company (within its duration in the index), we collect the corresponding earnings news within a time window ranging from the moment of the announcement up to the next market opening, that is, during the treatment window. As already mentioned, we only consider earnings announcements released outside the stock market opening hours.

- The ESS is a score ranging from 0 to 100 which represents the underlying sentiment for a news story and a given entity. A score of 50 indicates a neutral sentiment, while values above 50 indicate a positive sentiment and values below 50 a negative one. The ESS is determined using a proprietary algorithm, and it is derived from a collection of surveys where financial experts rate firm-specific events as conveying positive or negative sentiment and to what degree. We do not filter the data based on the ESS; however, as described above in the main text, we use the ESS to build our treatment variable.

Throughout our data analysis, we follow the Ravenpack ID (RP_ENTITY_ID) assigned by RPNA to uniquely identify each company in our dataset.

A.2 Financial outcomes and controls

To measure the outcomes and controls, we define the pre- and post-treatment windows, which include 8 calendar days (open to close), respectively. Note that if the first/last day in the pre/post treatment window is not a trading day, e.g., it happens when the announcement arrives on a Monday morning or a Friday evening, we additionally take the last/first 5 minutes on the previous/upcoming trading day into account. Any other holidays or unexpected data missing within the windows are ignored.

Three types of financial outcome variables are considered simultaneously in our analysis and measured over the post-treatment window: log-returns, realized volatilities, and liquidity. Log-returns and realized volatilities are computed using the high-frequency stock trading data obtained from the NYSE Trade and Quote (TAQ) database. The ticker symbols are matched with RP_ENTITY_ID according to the source mapping file provided by RPNA. The raw TAQ data are cleaned by following the customary steps, as suggested in Barndorff-Nielsen et al. (2009):

1. Only keep the transactions between 9:30-16:00 in each trading day (when the exchange is open) and with non-zero price.

2. Delete transactions with a correction indicator.

3. If multiple transactions have the same time stamp, replace all these with the median price.

We compute the daily log-returns by taking the log difference between the last and the first trade prices in the window 9:30-16:00. Prices are adjusted for dividend payments and stock splits. The realized volatility is computed using intraday log-returns $r_{t-i\cdot\Delta}$ estimated at a frequency Δ corresponding to the splitting of the time window 9:30-16:00 into N disjoint time subintervals. More specifically,

$$RV_t^{(d)} = \sqrt{\sum_{i=0}^{N-1} r_{t-i\cdot\Delta}^2}.$$

In our dataset, the frequency Δ will be set to five minutes. Finally, we shall proxy our last financial outcome variable, that is, the stock-specific liquidity, by using the daily turnover-ratio defined as the quotient trading volume (in number of shares)/shares outstanding. The relevant data are acquired from Compustat database. We aggregate the volatility and liquidity measures by averaging the daily estimators over the one-week window. Additionally, the three outcomes measured over the one-day pre/post-treatment windows are also used as responses to examine the short-run effects.

To control the impact of other economic and financial variables, we include global sector classification, NBER-based recession indicator, Fama-French (FF) 3 factors, CBOE Volatility Index (VIX), US Gross Domestic Product (GDP) and Consumer Price Index (CPI), which are observed over the *post*-treatment window, as control variables. In addition, FOMC meeting indicator, lagged outcomes, VIX, FF factors, and some relevant financial statement variables measured over the *pre*-treatment window are also included in the set of covariates. Note that for financial statement variables that are quarterly updated, we just take the last observation before earnings release. Table A.1 summarizes the detailed definition and data sources of the financial and macroeconomic controls considered in our analysis.

Data	Definition	Source
global sector classification	observation on the earnings announcement date	Compustat (quarterly updated)
NBER-based recession indicator	observation on the earnings announcement date	FRED (monthly updated)
FOMC meeting indicator	dummy over the pre-treatment window	FMOC
VIX	averaged daily observations over the window	CBOE
GDP (US, quarterly basis)	percent change from preceding period, seasonally adjusted annual rate	FRED
inflation measured by CPI (US, monthly basis)	annual growth rate (%)	OECD
Small (market capitalization) Minus Big	averaged daily observations over the window	Fama-French Portfolios and Factors
High (book-to-market ratio) Minus Low	averaged daily observations over the window	Fama-French Portfolios and Factors
excess return of the market portfolio	averaged daily observations over the window	Fama-French Portfolios and Factors
loading on the market risk	OLS estimates on the last calendar quarter before earnings release	Fama-French Portfolios and Factors, Compustat
log of total assets (quarterly updated)	last observation before earnings release	Compustat
market to book ratio (quarterly updated)	$\frac{\text{close price}\times\text{common shares outstanding}}{\text{stockholders equity}}$, last observation before earnings release	Compustat

Table A.1: Variable descriptions for the economic and financial controls

References

Andrei, D. and Hasler, M. (2015). Investor Attention and Stock Market Volatility. *The Review of Financial Studies* **28**(1), 33-72.

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.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science* **355**, 483-485.

Athey, S. and Imbens, G.W. (2016). Recursive Partitioning for Heterogeneous Causal Effects. *Proceedings of the National Academy of Sciences of the United States of America* **113**(27), 7353-7360.

Athey, S. and Imbens, G.W. (2017). The State of Applied Econometrics: Causality and Policy Evaluation. *Journal of Economic Perspectives* **31**(7), 3-32.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized Random Forests. *Annals of Statistics* **47**(2), 1148-1178.

Audrino, F., Sigrist, F., and Ballinari, D. (2020). The impact of sentiment and attention measures on stock market volatility. *International Journal of Forecasting* **36**, 334-357.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2009). Realized kernels in practice: Trades and quotes. *The Econometrics Journal* **12**(3), C1-C32.

Barberis, N., Shleifer, A., and Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics* **49**(3), 307-343.

Black, F. (1986). Noise. *The Journal of Finance* **41**(3), 528-543.

Christiansen, C., Schmeling, M., and Schrimpf, A. (2012) A comprehensive look at financial volatility prediction by economic variables. *Journal of Applied Econometrics*, **27**(6), 956-977.

Cookson, J. A., and Niessner, M. (2020). Why Don't We Agree? Evidence from a Social Network of Investors. *The Journal of Finance* **75**(1), 173-228.

Daniel, K., Hirshleifer, D., and Teoh, S. H. (2002). Investor psychology in capital markets: Evidence and policy implications. *Journal of monetary economics* **49**(1), 139-209.

Della Vigna, S. and Pollet, J. M. (2009). Investor inattention and Friday earnings announcements. *The Journal of Finance* **64**(2), 709-749.

De Long, J. B., Shleifer, A., Summers, L. H., and Waldmann, R. J. (1990). Noise trader risk in financial markets. *Journal of Political Economy* **98**(4), 703-738.

Engelberg, J., McLean, R. D., and Pontiff, J. (2018). Anomalies and News. *The Journal of Finance* **73**(5), 1971-2001.

Gargano, A. and Rossi, A. G. (2018). Does It Pay to Pay Attention? *The Review of Financial Studies* **31**(12), 4595-4649.

Giannini, R., Irvine, P., and Shu, T. (2019). The convergence and divergence of investors' opinions around earnings news: Evidence from a social network. *Journal of Financial Markets* **42**, 94-120.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd edition, Springer (10th printing with corrections, 2013).

Hirshleifer, D., Lim, S. S., and Teoh, S. H. (2009). Driven to Distraction: Extraneous Events and Underreaction to Earnings News. *The Journal of Finance* **64**(5), 2289-2325.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87**(3), 706-710.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences.* Cambridge University Press.

Imbens, G.W. and Wooldridge, J.M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* **47**(1), 5-86.

Johnson, E. J. and Tversky, A. (1983). Affect, generalization, and the perception of risk. *Journal of personality and social psychology* **45**(1), 20.

Kahneman, D. (1973). *Attention and effort.* Prentice-Hall series in experimental psychology.

Knaus, M. C., Lechner, M., and Strittmatter, A. (2020). Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence. *The Econometrics Journal*, forthcoming.

Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric Evaluation of Labour Market Policies*, Lechner, M. and Pfeiffer, E. (editors), 43-58. Physica.

Lechner, M. (2018). Modified Causal Forests for Estimating Heterogeneous Causal Effects. Working Paper, University of St. Gallen.

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

Mitra, G. and Mitra, L. (2011).. *The handbook of news analytics in finance*, volume 596. John Wiley & Sons.

Mittnik, S., Robinzonov, N., and Spindler, M. (2015). Stock market volatility: Identifying major drivers and the nature of their impact. *Journal of Banking and Finance*, **58**(Supplement C), 1-14.

Nonejad, N. (2017). Forecasting aggregate stock market volatility using financial and macroeconomic predictors: Which models forecast best, when and why? *Journal of Empirical Finance*, **42**(Supplement C), 131-154.

Paye, B. S. (2012) 'Deja vol': Predictive regressions for aggregate stock market volatility using macroeconomic variables. *Journal of Financial Economics*, **106**(3), 527-546.

Peng, L. (2005). Learning with information capacity constraints. *Journal of Financial and Quantitative Analysis* **40**, 307-330.

Peng, L., and Xiong, W. (2006). Investor attention, overconfidence and category learning. *Journal of Financial Economics* **80**(3), 563 - 602.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**(5), 688-701.

Siganos, A., Vagenas-Nanos, E., and Verwijmeren, P. (2017). Divergence of sentiment and stock market trading. *Journal of Banking & Finance* **78**, 130-141.

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.

Wager, S. and S. Athey (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Society* **113**(523), 1228-1242.

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