The kinks of financial journalism

Diego García*

Leeds School of Business
University of Colorado Boulder

May 7, 2018

ABSTRACT

This paper studies the content of financial news as a function of past market returns. As a proxy for media content we use positive and negative word counts from general financial news columns from the Wall Street Journal and the New York Times. Our empirical analysis allows us to discriminate between theories that predict hyping good stock performance to those that emphasize negative news. The evidence is conclusive: negative market returns taint the ink of typewriters, while positive returns barely do. Given how pervasive our estimates are across multiple time periods, subject to different competitive pressures in the market for news, we conclude our results are driven by demand considerations.

JEL classification: G01, G14.
Keywords: media content, stock returns, journalism.

---

*Leeds School of Business, University of Colorado Boulder, CB 419, Boulder, CO, 80309. Email: diego.garcia@colorado.edu; Webpage: http://leeds-faculty.colorado.edu/garcia/

1 Introduction

Financial markets have permeated investors' daily lives since the advent of newspapers. Columns with business news, with commodity and stock price movements and commentary around them, have been a staple in journalism virtually since the invention of the printing press.[1] This paper studies the content of a large set of columns on general finance news from the New York Times and the Wall Street Journal for the time period 1905–2005. These columns covered general financial news, from general stock market trends, to macroeconomic events and discussions on individual companies. They were published every day throughout our sample, generating a textual corpus that is quite homogeneous, and, as such, is a nice laboratory where to study how journalists colored different economic events for their readers.

Our empirical approach models media content by the fraction of positive and negative words used by journalists as a function of market returns over the recent few days. The returns of the Dow Jones Industrial Average (DJIA) are excellent predictors of journalists' word choices in our columns: higher market returns lead to journalists using more positive words, and less negative words (Tetlock, 2007). A parsimonious model with only lagged market returns can explain more than 30% of the variation in media content. More importantly, the effect is highly non-linear: positive returns have a much smaller impact on content than negative returns. This effect is even more pronounced for lagged returns that occurred days before the writing of the article: past returns only influence media content when they are negative.

Our results suggest that the "hyping" in Shiller (2000) cannot generally be associated with high market performance. On the contrary, it is the negative domain that seems to excite the imagination of journalists. The evidence is consistent with models of human behavior where the domain of gains and losses trigger different reactions by agents. This asymmetry corroborates one of the main tenants of behavioral economics, the loss-aversion of Kahneman and Tversky (1979)'s prospect theory, as well as laboratory studies on the different perceived stimuli in the

[1] For example, the "Notizie scritte," first published by the government of Venezia in 1556, or "Relation," released in 1605, both included economic news. Daily newspapers, very close to what we have today, were published widely in the United Kingdom and continental Europe by the 18th century. The Economist was founded in 1843, whereas The Financial Times and The Wall Street Journal were founded in 1888 and 1889 respectively.

domain of gains versus losses (Kuhnen, 2015).

The competition in the market for financial news was quite limited during most of the 20th century. Most investors only had access to a few sources for information on Wall Street, among them the two leading newspapers we study. But our sample includes the period starting in the 1980s, where CNN, followed by CNBC, and later the Internet cut into the turf of the more traditional print media. Thus, we can exploit time-series variation in the supply side of the market to see how competition affected the way journalists slanted financial news for investors. We show the non-linearities we uncover are pervasive throughout our sample period. The evidence from the later 25 years of our sample (1980–2005), where the supply of information increased significantly, exhibits the same pattern as in the 1920s (or the 1950s). Furthermore, even when conditioning on the author of the column under consideration, we find consistent emphasis on negative news. Given the larger changes in the media landscape that occurred throughout sample, and the idiosyncrasies of authors' writing styles,[2] we conclude that our results are more likely to be driven by demand-side considerations (Mullainathan and Shleifer, 2005).

One can consider our empirical study a nice laboratory where to look at impulse-response "in the field." In contrast to other type of news (i.e., politics), the sources of discussion in our articles (stock price movements, earnings, macroeconomic announcements) are easily measurable, and by and large, they are continuous variables, i.e. DJIA returns. The columns we study are colored around such variables, which gives us a particularly sharp test of the impulse response function from stimuli (market conditions) to the printed page. The sources we include in our analysis were the leading media providers of financial news in the United States for the majority of our sample period, which allows us to study how journalists chose to slant the facts from financial markets for their readers.

Most of the literature studying the media in economics has focused on political news, and in particular their bias towards Republicans or Democrats.[3] For example, Gentzkow and Shapiro

[2]Dougal, Engelberg, García, and Parsons (2012) document large differences in writing style, in terms of words-per-sentence, complexity of the text, as well as unconditional pessimism.

[3]See DellaVigna and Gentzkow (2010) for a survey on empirical studies on persuasive communication, which discusses literature from marketing, psychology and sociology as well. The survey by Prat and Strömberg (2011)

(2010) construct a metric of media slant based on the language used by media outlets, and argue that reader's preferences in the political spectrum are the key drivers of the newspapers' content.[4] Our paper contributes to this literature by considering a very different type of slanting, in which journalists have a continuous (multi-dimensional) variable to convey to their readers, rather than a dichotomous (left-right) choice. As in politics, our evidence suggests demand-side considerations are more important than supply-side variation.

Starting with Shiller (2000), financial economists have studied the effect of the media on equilibrium outcomes. The focus of these papers is on the effect of the media on asset prices (Tetlock, 2007; Fang and Peress, 2009), trading behavior (Barber and Odean, 2008; Engelberg and Parsons, 2011), corporate governance (Dyck, Volchkova, and Zingales, 2008), merger negotiations (Ahern and Sosyura, 2014), and IPO returns (Hanley and Hoberg, 2012; Loughran and McDonald, 2013).[5] In contrast, our paper focuses on the drivers of the media content itself, i.e., on the effect of asset prices and corporate events on what journalists choose to write about. A subset of these papers study predictors of media content mostly in order to create instruments to argue causality in a second stage.[6] Finally, there is another literature that argues how advertising revenues can affect the slant and coverage of the media. For example, Reuter and Zitzewitz (2006) show how past advertising influences mutual fund recommendations in personal finance publications. Ellman and Germano (2009) study how such economic ties can generate biases even in the presence of competition.

In the journalism literature, there are several studies of content as a function of economic

also discusses theoretical contributions to the literature.

[4]The literature is rapidly growing, see Groseclose and Milyo (2005), Gentzkow and Shapiro (2006, 2008, 2011), Gentzkow, Shapiro, and Sinkinson (2011), Baum and Groeling (2008), Iyengar and Hahn (2009), Enikolopov, Petrova, and Zhuravskaya (2011), Larcinese, Puglisi, and Jr. (2011), Chiang and Knight (2011) for some recent examples.

[5]See Cutler, Poterba, and Summers (1989), Klibanoff, Lamont, and Wizman (1998), Huberman and Regev (2001), Chan (2003), Dyck and Zingales (2003), Das and Chen (2007), Gaa (2008), Tetlock (2010), Tetlock, Saar-Tsechansky, and Macskassy (2008), Engelberg (2008), Bhattacharya, Galpin, Yu, and Ray (2009), Tetlock (2010), Griffin, Hirschey, and Kelly (2011), Solomon (2012), García (2013), Jegadeesh and Wu (2013), Hillert, Jacobs, and Müller (2014), Neuhierl, Scherbina, and Schlusche (2013), Kogan, Routledge, Sagi, and Smith (2015), Heston and Sinha (2016), Mamaysky and Glasserman (2017), Boudoukh, Feldman, Kogan, and Richardson (2016), Manela and Moreira (2017) for other research on the effect of media in the finance literature. Loughran and McDonald (2016) surveys the literature with an emphasis on accounting topics.

[6]For example, Engelberg and Parsons (2011) and Gurun and Butler (2012) use geography to predict content, Dougal, Engelberg, García, and Parsons (2012) use author's identity, Peress (2014) uses newspapers strikes for identification.

variables. Bow (1980) argues that there were no predictive signs in the media prior to the 1929 stock market crash, while Griggs (1963) gives a similar account in the context of the 1957–1958 recession. Neilson (1973) discusses the state of journalism during the bulk of our sample. Norris and Bockelmann (2000) and Roush (2006) also have extensive discussions as to the role of the media since the beginning of the 20th century.

Also related is the literature on financial advice more generally. For example, Barber and Odean (2008) show how retail investors tend to favor stocks that are in the news, whereas Foerster, Linnainmaa, Melzer, and Previtero (2015) focus on how financial advisors exert substantial influence over their clients' portfolios, presumably via the advice they directly provide. While our study does not speak to actual investment choices, we get to read widely circulated financial news which are intended to inform their readers, presumably to help them with their own investment decisions.

The rest of the paper is structured as follows. In Section 2 we introduce the data we use in our empirical analysis. Section 3 discusses our main results, while Section 4 looks at author fixed effects, different time periods, lower-frequency returns, as well as other indexes and different measures of media content. Section 5 concludes.

2 The data

The paper uses several sources of data. The first is stock return information. For the majority of our analysis, we use the Dow Jones Industrial Average from Williamson (2008).[7] The Dow Jones Industrial index goes back to the turn of the 20th century, and thus allows us to have a metric of US stock returns prior to the coverage in the more standard Center for Research in Security Prices (CRSP), which started in 1926. We let R_t denote the log-return on the DJIA index on date t. Business cycle information is obtained from the NBER website `http://www.nber.org/cycles.html`. The main source of data is the media content of three

[7] Historical data is available from `http://www.djaverages.com/`, including the total return for the Dow Jones Industrial Average, but this source does not include Saturday data. For this reason, we use the data on the DJIA from Williamson (2008), see `http://www.measuringworth.org/DJA/`. Exclusion of the Saturday data does not affect any of our results.

different columns of financial news, which we describe next.

The media content measures are constructed starting from the Historical New York Times Archive and the Historical Wall Street Journal Archive. The former goes back to the origins of the newspaper in 1851, while the latter starts in 1889. These datasets were built by scanning the full content of the newspaper, cropping columns separately. In order to have a consistent set of articles that cover financial news, we focus on three columns that were published daily during this period. From the New York Times we use the "Financial Markets" column, and the "Topics in Wall-Street" column (García, 2013), and from the Wall Street Journal we use the "Abreast of the market" column (Tetlock, 2007). The "Topics in Wall-Street" column ran daily under different titles (i.e. "Sidelights from Wall-Street", "Financial and Business Sidelights of the Day," "Market Place") until the end of our sample period. The "Financial Markets" column stopped being published with such a heading in the 1950s, although the New York Times obviously continued to publish a column with the financial news for the day, which we use in our analysis. The "Abreast of the Market" column was published daily virtually uninterrupted from 1926–2007, see Tetlock (2007) for details. This paper studies a total of 76,537 pdf files from the Historical Archives that were associated with either of these columns from January 1, 1905 through December 31, 2005.[8] A total of 55,168 of the columns in our sample were from the New York Times, while the other 21,369 are from the Wall Street Journal.

The columns under study were essentially summaries of the events in Wall Street during the previous trading day. The average article had around 800 words. The articles discussed anything from particular companies or industries to commodities and general market conditions. The topics included in the columns were of a business nature, with a focus on financial matters. Tetlock (2007) and García (2013) give more detailed accounts of the data sources.

To construct the media content measures, we transform the scanned images available from the New York Times Historical Archive into text documents. This is referred to in the computer science literature as "optical character recognition" (OCR). We use ABBYY, one of the leading software packages in OCR processing, to convert the images into text files. Although the quality

[8] We exclude news from the period in 1914 when the New York Stock Exchange was closed (through December 12th, 1914).

of the transcription of the articles is high, it is important to notice that the accuracy of OCR processing may be low for some files, particularly early in our sample. The quality of the scanned images in the NYT Historical Archive is low prior to 1905, thus our choice of starting date.[9] We note that this approach to reading text only adds random noise to our media content measures, and thus it will not bias our conclusions.

In order to quantify the content of the New York Times articles, this paper takes a "dictionary approach."[10] For each column i written on date t, we count the number of positive words, g_{it}, and negative words, b_{it}, using the word dictionaries provided by Bill McDonald.[11] As argued in Loughran and McDonald (2011), standard dictionaries fail to account for the nuances of Finance jargon, thus the categorization we use has particular merits for processing articles on financial events. We let w_{it} denote the total number of words in an article. We construct these media measures dating them to the day t in which they were written, with the understanding that they are published in the morning of day $t + 1$. The rationale is that the information contained in these columns clearly belongs to date t. The writing process for each article started at 2:30-3:00 pm, typically just as the market was about to close, and the final copy was turned in to be edited and typeset at around 5-6 pm.

We aggregate the media content measures to create a time-series that matches the Dow Jones index return data available. In particular, we first combine all news printed between two trading dates, in order to be conservative with our standard errors, and also reduce noise stemming from particular idiosyncrasies from each column. In essence, we are trying to measure the content of the financial news on investors' desks prior to the opening of the market, and model its relationship to previous market events. We will also study cross-sectional variation with respect to each column, since the Sunday and Monday columns are likely rather different, both in terms of circulation and material on which to write about, from the other weekday columns.

[9]The OCR software will try to interpret anything in the original image, from spots to actual text. Different margins, multiple columns, and page formatting issues in general present a challenge for the character recognition process.

[10]Non-dictionary approaches have gained much popularity in recent research on text content analysis, in which not just the words, but the order and their role in a sentence is taken into account (see Jurafsky and Martin, 2018, for details). Given the OCR processing issues discussed above, these types of language processing algorithms are more challenging to implement in our study.

[11]See `http://www.nd.edu/~mcdonald/Word_Lists.html` for details.

In order to aggregate the news, we average the measures of positive/negative content from articles that were written since the market closed until the market next opens. When the market is open on consecutive days, t and $t+1$, we define our daily measure of positive media content as $G_t = \sum_i g_{it} / \sum_i w_{it}$, where the summation is over all articles written in date t (given our news selection, there are at least two such articles for the vast majority of days in our sample). Similarly, we construct our daily measure of negative media content as $B_t = \sum_i b_{it} / \sum_i w_{it}$. In essence, we count the number of positive and negative words in the financial news under consideration, and normalize them by the total number of words. For non-consecutive market days we follow a similar approach, including all articles published from close to open. To be precise, consider two trading days t and $t+h+1$ such that $h > 0$ and the market was closed h days, from $t+1$ through $t+h$. We define the positive media content measure as $G_t = \sum_{i,s=t}^{s=t+h} g_{is} / \sum_{i,s=t}^{s=t+h} w_{sh}$. We proceed analogously for the negative media content variable and define $B_t = \sum_{i,s=t}^{s=t+h} b_{is} / \sum_{i,s=t}^{s=t+h} w_{sh}$. We define media content as the difference between the positive and negative media content measures, i.e. $M_t = G_t - B_t$.

For consecutive trading dates, our media measures G_t and B_t are constructed using information that was available as of the end of date t when the market is open on date $t+1$ (the bulk of our sample). It is less clear whether market prices on date t reflected the information available to the journalists writing the columns, as the deadline for turning in the article to the editor was not until roughly 5-6pm, while the NYSE closed at 3–4pm. We further remark that for non-consecutive trading dates, we use articles that may have been written on days after date t, but prior to the market opening (i.e., in the case of holidays).

Table 1 presents summary statistics on our media measures. The average number of positive and negative words, averaged over all articles, are 1.27 and 2.08 respectively. The articles from the Wall Street Journal use slightly higher fraction of positive words, 1.42 versus 1.21, but the fraction of negative words are virtually identical, 2.08 versus 2.09. We remark how our time-series aggregate, which adds all articles between trading dates, have similar means. The standard deviation of this time series aggregate is significantly less noisy. For example, positive content has a standard deviation from 0.36 when aggregated, versus 0.63 when looking at individual articles. The media content variable, which simply subtracts negative from positive frequencies, inherits the properties just discussed for their individual components.

For the rest of the paper we normalize our sentiment measures so they have zero mean and unit variance. This will allow us to interpret the regression coefficients in terms of one-standard deviation shocks to the sentiment measures, thus making it easier to gauge the economic magnitude of our results.

3 Media content and DJIA returns

The build our main results in three steps. We first document in Section 3.1 the strong relationship between stock returns and media content, corraborating the previous findings in the literature (Tetlock, 2007; García, 2013), in particular their "long-memory." In Section 3.2 we present the main results of the paper, documenting different slanting by financial journalists of positive versus negative market returns. In Section 3.3 we study the effect of lagged returns on media content, focusing again on the differential effect of stock returns in the positive/negative domains.

3.1 Linear models

We start by estimating a parsimonious time-series model of media content. In particular, we assume the following econometric specification:

$$M_t = \beta_0 R_t + \beta \mathcal{L}(R_t) + \rho \mathcal{L}(M_t) + \eta X_t + \epsilon_t; \tag{1}$$

where M_t denotes the media content written between trading dates t and $t+1$ (for articles written after the market closed on date t by prior to opening on date $t+1$), and R_t denotes the log-return on the DJIA on date t (from close on date $t-1$ to close on date t). We truncate R_t at -3% from below and $+3\%$ from above in the analysis that follows. The set of explanatory variables X_t includes day-of-the-week dummies and a cubic function of time.

Table 2 presents estimates of (1). The first set of columns ("Only $\mathcal{L}(R_t)$") present estimates under the constraint $\rho = 0$, while the second set of columns ("Only $\mathcal{L}(M_t)$") present estimates under the constraint $\beta = 0$. The last two columns present the unconstrained estimates. The t-statistics reported in the table use Newey-West corrections with ten lags.

The most important determinant of media content is the last trading day returns, as expected. A one-standard deviation shock to returns moves media content by one-half of a standard deviation, a large effect in economic terms. The statistical significance is also very large, due to the large sample we are studying, and the strong correlation between M_t and R_t. Perhaps more surprising is the fact that lagged returns, even nine days into the past, have significant predictive power. The economic magnitudes decline quickly with the distance between the return date and the writing date, but the aggregate effect of returns lagged 5–9 days is non-trivial. Overall, lagged returns explain 34.6% of the variation in media content.

The second column presents the estimates ignoring lagged returns, but including lagged media content. The autocorrelation structure given in columns 4-5 of Table 2 shows that M_t is a fairly persistent process. But lags of media content can only explain 22.4% of the variation in media content itself. The last two columns give the unconstrained model. We highlight how the introduction of lagged media content does increase the R^2 of the regression, from 34.6% to 42.3%. The autocorrelation of media content is 0.141 in this specification, which suggests there is a persistent component, but it is not very large. This persistence can be easily explained by author fixed effects, for example, as the same journalists would write the columns at hand during different periods of time, each with their own style (Dougal, Engelberg, García, and Parsons, 2012).

It is important to note that the coefficient on R_t is still virtually unchanged, 0.478 versus 0.470. Furthermore, the impulse response of M_t to R_{t-k} is also not different from that implied by the first set of columns to the last set. For example, the impulse response to R_{t-1} is $\beta_1 = 0.149$ in the first specification, and $\beta_1 + \rho_1\beta_0 = 0.143$ in the last. The impulse response to R_{t-2} is $\beta_2 = 0.082$ in the first specification, and $\beta_2 + \rho_1\beta_1 + \rho_1^2\beta_0 + \rho_2\beta_0 = 0.076$ in the last.

3.2 The main kink

Our main specification to capture non-linearities will consist of a model of media content of the form:

$$M_t = f(R_t; \alpha, \beta) + \eta X_t + \epsilon_t; \tag{2}$$

where M_t denotes the media content written between trading dates t and $t+1$ (for articles written after the market closed on date t by prior to opening on date $t+1$), and R_t denotes the log-return on the DJIA (from close of date $t-1$ to close of date t), truncated at -3% from below and $+3\%$ from above. The set of explanatory variables X_t includes 10 lags of R_t and M_t, as well as day-of-the-week dummies and a cubic function of time.

Since our main interest is on potential non-linearities between the outcome variable, media content, and stock returns, we propose a parsimonious, yet flexible function $f(R_t; \alpha, \beta)$. In particular, we assume that the function f is of the piecewise linear form

$$f(R_t; \alpha, \beta) = \sum_{i=1}^{4} (\alpha_i + \beta_i R_t) 1_{\{R_t \in S_i\}} \tag{3}$$

where the sets S_i are: $S_1 = [-3, -1)$, $S_2 = (-1, 0)$, $S_3 = (0, 1)$, and $S_4 = (1, 3]$.

Our discussion will focus on the slope parameters β_i, in particular we will compare their magnitudes in the "tails," β_1 versus β_4, as well as in the "middle," β_2 versus β_3. Our main tests will try to reject the null that such slopes are the same. For completeness we include in our tables the estimated α_i's, and we will also test for differences between α_2 and α_3 in our tables, since rejecting the null of equality of such coefficients implies the existence of a jump at zero, a natural reference point.

We also estimate a model with "smooth" non-linearities, which we use in our plots. While such a model cannot estimate the jump at zero, it does corroborate our main parametric conclusions.[12] The choice of the set of intervals S_i is actually motivated by the fit of a model based on splines. The fact that we impose a linearity restriction for each interval makes hypothesis

[12]Estimating the model with splines that allow for a discontinuity at zero only reinforces our findings with respect to a jump at zero.

testing and the interpretation of coefficients simply more transparent.

Table 3 presents the estimates of the model in (2). We focus first on the slope coefficients, β_i, given in the right. Not surprisingly, the content of financial news is in large part determined by the market returns during the last trading day (proxied by the DJIA). All parameters β_i, $i = 1, \ldots, 4$, are significant at standard levels of significance. Two differences stand out: (1) the slope in the negative "normal" range $(-1, 0)$ is $\beta_2 = 0.656$, whereas that in the "normal" positive range $(0, 1)$ is only $\beta_1 = 0.528$. While both are significantly higher than the slopes at "tail" market returns (above 1% or below -1%), the slope on the negative domain is significantly higher than on the positive domain. The test on Panel B yields a Newey-West adjusted test with a p-value well below 1%.[13]

When the DJIA rose by more than 1%, the news content becomes significantly less sensitive to market returns. The point estimate of $\beta_4 = 0.057$ is statistically significant, but small in economic terms. Moving DJIA returns from +1% to +3% changes the content of financial news by little more than 1/10th of a standard deviation. In contrast, in the negative "tail" domain, the coefficient $\beta_1 = 0.203$ tells us a similar market move would change media content by 4/10ths of a standard deviation. As the test in the last row of Panel B documents, the difference is significant.

The tests of the slopes just discussed, both around zero and for larger market moves, shed some light on the differences in writing on the domain of gains and losses. We have rejected the null hypothesis of a function that has the same slant in the positive and negative domain. Our next test studies whether there is a jump in the news content itself around the natural reference point of zero returns, that is, whether the function f is discontinuous at zero. Given our parametric specification, we can test for such a jump simply by comparing the intercept coefficients α_2 and α_3. Their difference is 0.152, which is non-trivial in economic terms, and highly statistically significant, as reported in Panel B. We conclude that there is indeed a difference between reporting very small positive returns, versus reporting very small negative returns. The boundary of the domain of gains and losses acts as a reference point for journalists.

[13] All statistics reported use Newey-West corrections with ten lags.

3.3 Kinks with recent past returns

In our next set of tests, we augment the specification in (2) by adding non-linear functions of DJIA returns two to four days before the publishing of the papers (measured on trading days time). This is a nice "out-of-sample" test of our previous results, since both Tetlock (2007) and García (2013) document that media content is influenced by lags going back at least four trading periods.[14] While there is little reason to suspect that the returns 2–4 days ago would have much of an effect as a "reference point," the point estimates on the slope coefficients act as an alternative test of the "kinks" documented in Table 3 and Figure 1.

In particular, we estimate the model

$$M_t = f_1(R_t; \alpha_1, \beta_1) + f_2(R_{t-1}; \alpha_2, \beta_2) + f_3(R_{t-2}; \alpha_3, \beta_3) + f_4(R_{t-3}; \alpha_4, \beta_4) + \eta X_t + \epsilon_t;$$

where M_t denotes the media content written between trading dates t and $t+1$ (for articles written after the market closed on date t by prior to opening on date $t+1$), R_t denotes the log-return on the DJIA (truncated at -3% from below and $+3\%$ from above). The set of explanatory variables X_t includes day-of-the-week dummies and a cubic function of time.[15] The functions $f_j(R_t; \alpha, \beta)$ are assumed to be of the form

$$f_j(R_{t-j}; \alpha_j, \beta_j) = \sum_{i=1}^{4} (\alpha_{ji} + \beta_{ji} R_{t-j}) 1_{\{R_{t-j} \in S_i\}} \tag{4}$$

where the sets S_i are: $S_1 = [-3, -1)$, $S_2 = (-1, 0)$, $S_3 = (0, 1)$, and $S_4 = (1, 3]$. In essence, we reproduce the results from the main specification allowing for non-linearities for all lagged returns R_t through R_{t-3}. As before, all statistics reported in the table use Newey-West corrections with ten lags.

Panel A.1 in Table 4 mimics the first panel in Table 3, showing the relationship between

[14] Other low-frequency variables, such as GDP, also influence the writing, but at the daily frequency we are working with, their influence is both economically and statistically small.

[15] We do not include $\mathcal{L}(M_t)$ in (3.3) in order to avoid computing impulse response functions using our non-linear specification, a non-trivial task. The results in Table 2, discussed at the beginning of this section, suggest such omission does not bias our conclusions.

media content and the previous day DJIA returns, controlling for non-linearities with respect to lagged returns R_{t-2} through R_{t-4}. The estimated differences between the slopes in the positive and negative domain are slightly larger than in the previous specification: around zero the estimates now are 0.699 and 0.516, compared to 0.656 and 0.528 before; whereas in the tails the slopes are 0.240 and 0.027 now, versus 0.203 and 0.057 before. Thus the previous results are not affected by a slightly different econometric specification.

Looking at the slope coefficients for lagged returns in Panel A.2, we see that media content increases with stock returns two trading days ago, but exclusively in the negative domain (the sets S_1 and S_2). Remarkably, this result, which showed strongly in Table 3, comes out as strong in the regressions with the DJIA returns lagged two through four days. This is apparent in Panel A.2, where the point estimates for the slopes around zero are 0.248 (negative domain) and 0.109 (positive domain), compared to 0.699 and 0.516 for the one-day lag. While they are smaller, as expected, their difference is large in economic terms. The same pattern emerges for 3-days and 4-days lagged returns, the slopes around zero are 0.155 and 0.099 (negative domain) and 0.047 and 0.006 (positive domain).

While the statistical power at the "tails" is smaller, Panel A.2 of Table 4 documents very strong differences in the intervals $(-3, -1)$ and $(1, 3)$ for the returns from two days ago. The slope coefficient in the negative domain is 0.185, versus only 0.025 in the positive domain. The slopes at the "tails" for the returns three and four days before writing exhibit a similar pattern: the slope in the negative domain is positive and statistically different from zero at standard levels of confidence, whereas the slope in the positive domain is actually negative.

Panel B of Table 4 reports formal F-tests of one-dimensional restrictions, as in Table 3. All claims in the previous discussion hold. With respect to slope tests, all of them are highly significant with the exception of the tests with returns four days ago (highest p-value 2%). We conclude that there is no effect of lagged positive returns on the media content, whereas lagged negative returns do affect journalists' word choices.

4 Ancilliary results

In the previous section we established that the response of media content to lagged market returns was more pronounced in the negative domain than in the positive domain. In this section, we explore what may drive the kinks reported in Section 3. We first look at variation of the authorship of columns, and also analyze to what extent the evidence provided in Table 3 is stable throughout our sample period. We also extend our analysis by looking at a value-weighted index, as well as indexes of small and large firms.[16] We then ask whether the asymmetries may have a time-series component by looking at variation along the business cycle. We also study to what exteant the parametric assumptions, in particular the functional form (3), affect our inferences. We finally consider other media metrics, in particular contructing our sentiment measure using tf-idf weights.

4.1 Author fixed-effects

We start by studying to what extent the particular type of slanting we uncovered in Section 3 varies with the author of the column. Dougal, Engelberg, García, and Parsons (2012) document significant variation among authors regarding writing style, so there are reasons to suspect that who writes (supplies) the news will matter for the coloring of financial events. We use the authorship data from Dougal, Engelberg, García, and Parsons (2012), which covers the 1970-2005 period, in our next set of tests.[17] Table 5 presents estimates of the main specification in (2), estimated separately for the ten most prolific authors of the "Abreast-of-the-market" column from the Wall Street Journal.

The point estimates for Hillery, who wrote over 2,413 columns in our dataset, are given in the first row. Compared to the evidence in Table 3, it is clear that Hillery was not very different from the average column throughout our sample: the slopes over the negative domain are $\beta_1 = 0.333$ and $\beta_2 = 0.871$, significantly higher than those over the positive domain, $\beta_4 = 0.113$

[16] The indexes are from Ken French's data library. We note that the DJIA index used in our main analysis consists of only thirty firms.

[17] Journalists started signing their articles in the published versions of the Wall Street Journal and the New York Times around 1970.

and $\beta_3 = 0.636$. Comparing the magnitudes across the ten authors, we see that β_2 is larger than β_3 in eight out of ten cases. While the effects in the tails, β_1 and β_4, are estimated with much more noise, it is worthwhile highlighting how six out of the ten β_4 estimates are actually negative, which lines up with the previous evidence of the small effect of large positive returns on the slanting of the journalists.

4.2 Time series and business cycle variation

Table 6 replicates the analysis in Table 3, for four different time periods. If our findings were driven by supply side considerations (i.e. journalists peculiarities, competition among media providers), we should find different estimates through our sample period. We note that both the editor of the financial section, as well as the team of journalists writing the stories, changed multiple times during our sample.[18] Furthermore, as discussed in the introduction, the competitive landscape changed dramatically with the proliferation of cable TV in the 1980s, and the Internet in the 1990s.

The evidence in Table 6 suggests that the non-linearities previously reported are prevalent throughout our sample period. The estimates for β_1, for example, which measures the slope in the left-tail of market returns, are 0.241, 0.255, 0.286, and 0.236, whereas those in the right-tail, measured by β_4, are 0.024, 0.067, 0.094, and 0.116, for the time periods 1905–1930, 1931–1955, 1956–1980, and 1981–2005 respectively. It is rather remarkable how stable the non-linearities from Table 3 are: in all four time periods we have that $\beta_1 > \beta_4$, and that $\beta_2 > \beta_3$, i.e. the reaction of news to market returns are significantly higher in the negative domain.

Our next set of tests asks whether the impulse responses we have documented vary along the business cycle. There are reasons to suspect that this may be the case, from marginal utility arguments to theories based on psychology and mood during good and bad times (García, 2013). Table 8 estimates two different non-linear functions, with the parametric representation in (3), one during expansions, one during recessions, using NBER definitions. Columns 2–3 in Table 8

[18]See Dougal, Engelberg, García, and Parsons (2012) for a study of the role of journalists in financial news. In particular, they document significant variation in the rotations of journalists writing the "Abreast-of-the-market" column that we study.

present the results on expansions, whereas columns 4–5 include those for recessions. The non-linear fit from our main specification in Table 3 almost gets copied across the two sets of columns: there is little reaction to large positive returns, in contrast to the effect of large negative market movements. Even the magnitudes are very similar in economic terms. We conclude that the kinks that this paper documents are a general pattern that does not hinge on particular market states.[19]

4.3 Other indexes and intervals

Table 7 replicates the analysis in Table 3, using three different indexes. Our previous results are virtually unchanged, if anything slightly larger in magnitude. The models all detect a jump at zero. Furthermore, the tail slopes on the positive domain are all below 0.06, and statistically insignificant. This is in contrast to the point estimates on the negative domain, which are all large, from 0.14 to 0.28, and statistically significant. There is evidence that large stocks matter more in terms of media content, as expected, but small stocks returns also help predict media content, with similar kinks to those reported in Table 3.

There were three tests conducted in Table 3: two tests regarding behavior in the "middle" ($\alpha_2 = \alpha_3$ and $\beta_2 = \beta_3$), and one on slopes at the tails ($\beta_1 = \beta_4$). We mimic the analysis next by extending the number of linear splines. In particular, assume that the function $f(R_t; \alpha, \beta)$ is of the form

$$f(R_t; \alpha, \beta) = \sum_{i=1}^{8} (\alpha_i + \beta_i R_t) 1_{\{R_t \in S_i\}} \tag{5}$$

where the sets S_i are: $S_1 = [-3, -2)$, $S_2 = (-2, -1)$, $S_3 = (-1, -0.5)$, $S_4 = (-0.5, 0)$, $S_5 = (0, 0.5)$, $S_6 = (0.5, 1)$, $S_7 = (1, 2)$ and $S_8 = (2, 3]$.

We start discussing the tests on the tails of market returns. Looking at the columns labelled "Slopes," we find even stronger evidence of asymmetries in Table 9, compared to Table 3. Both β_1 and β_2 are positive, and significantly different than zero. On the other hand, β_7 is barely positive, and β_8 is actually negative. The fact that there is no "hyping" of large positive returns

[19] The NBER business cycle dummies are a good proxy for most other "market downturn" proxies one can empirically develop. In unreported results we find the non-linearities we document in Table 3 in time-series subsets sorted by lagged market returns and market volatility.

is very robust, as is the fact that large negative returns do influence journalists' word choices.

Turning to the behavior around zero market returns, we find that, in contrast with the results in Table 3, the slope around zero is slightly higher in the domain of gains than that of losses, $\beta_5 > \beta_4$, albeit the difference is not statistically significant (p-value 0.271). There is also no slope difference in the $(-1, -0.5)$ and the $(0.5, 1)$ ranges. In contrast, the model still detects a significant jump at zero.

These findings hold across a large set of non-linear models. The slope on the positive domain just above the zero market returns point is rather steep (see also Figure 1). Inferences just around zero on slopes are unconclusive. But starting at -0.5 and $+0.5$ these differences start to be noticeable, and they become more pronounced as we move out in the gains and losses domains. The differences between the behavior in the right tail and the left tail, for any reasonable definition of tails, are both statistically and economically significant.

4.4 Other media metrics

In our last battery of tests, we consider different approaches to measure content. In our main analysis we have used the bag-of-words approach of Tetlock (2007), specialized to the dictionaries developed by Loughran and McDonald (2011). In untabulated results, we use the Harvard-IV dictionary originally studied in Tetlock (2007), and our results are both qualitative and quantitatively unaltered.

A potential problem with the dictionary approach is the fact that some words may appear more often than others for reasons beyond the "sentiment" we are trying to measure. For example, top positive words, in terms of frequencies, are "gain" and "advance," whereas top negative words include "drop," and "decline," which most likely are included as boilerplate language to describe positive/negative market movements. The computer science literature has tackled this problem by introducing the concept of tf-idf weights (term frequency-inverse document frequency), via which words that occur more frequently are weighted down, whereas more infrequent words are given higher weights.

Figure 3 presents the estimates using tf-idf weights when measuring media content. The media content metric is now "noisier," since we are removing many of the purely descriptive words. But there is still a clear strong relationship between the previous day DJIA returns and the tf-idf weighted content in the negative domain. The magnitude of the slopes is similar to those in Figure 1, if anything the slopes are steeper in the left-tail of the return distribution. More importantly, the slope in the positive domain is virtually zero: it even shows a (non-significant) negative slope in the right-tail.

One last concern, and a natural linguistic question, is the different role that positive and negative words can play in our results. Figure 4 mimics our previous empirical analysis, using tf-idf weights, separately for positive and negative words. The left-panel shows how positive words follow a very similar pattern to that in Figure 1: while journalists do not use more positive words as returns move from 0.5–3%, they are using a lot less positive words when moving from 0 to -2%. The right panel, which presents the estimates for negative words, presents a similar picture: very steep slope in the negative domain, and flat/positively-sloped in the positive domain. Borrowing from John Authers' discussion of our paper:[20] "[...] journalists are far more scared of encouraging readers to buy and ushering them into a loss, than we are to be cautious, and leading them to miss out on a gain." Using more negative words when the market is going up may be a way to advice investors to be prudent with their gains.

5 Conclusion

This paper has established a strong non-linearity between lagged market returns and the content of financial news. The shape of the relationship is present in all subsamples we have studied, and holds not just for the last trading day, but for returns 2–5 days before publication. The evidence is very stable throughout our sample period, both in terms of the actual decade when the news were written, or the economic conditions (recessions/expansions), as well as who actually wrote the news.

We conclude that a demand-driven theory is a more plausible description of our data: in-

[20] John Authers is a regular contributor to the Financial Times, writing columns very similar to those studied here. See `https://www.ft.com/content/c884defa-054a-11e7-ace0-1ce02ef0def9`

vestors want the journalists to color the financial news emphasizing the negative domain. But this interpretation is indirect, as it relies on news supply changes affecting how journalists compete with each other, with the implicit assumption that such changes will affect how they write. While we can document strong kinks in how newspapers color market movements, our study cannot answer why investors, on average long the US market, want to get their financial news slanted after a bad day in Wall Street, but not after positive market movements. Further research is necessary to get at the mechanism behind the new empirical facts document in our paper.

References

Ahern, K. R., and D. Sosyura, 2014, "Who writes the news? Corporate press releases during merger negotiations," *The Journal of Finance*, 69(1), 241–291.

Barber, B., and T. Odean, 2008, "All that glitters: the effect of attention and news on the buying behavior of individual and institutional investors," *Review of Financial Studies*, 21(2), 785–818.

Baum, M. A., and T. Groeling, 2008, "New media and the polarization of American political discourse," *Political Communication*, 25(4), 345–365.

Bhattacharya, U., N. Galpin, X. Yu, and R. Ray, 2009, "The role of the media in the internet IPO bubble," *Journal of Financial and Quantitative Analysis*, 44, 657–682.

Boudoukh, J., R. Feldman, S. Kogan, and M. Richardson, 2016, "Which News Moves Stock Prices?," working paper, Working paper, IDC.

Bow, J., 1980, "The "Times's" Financial Markets column in the period around the 1929 crash," *Journalism Quarterly*, 57, 447–450.

Chan, W. S., 2003, "Stock price reaction to news and no-news: drift and reversal after headlines," *Journal of Financial Economics*, 70, 223–260.

Chiang, C.-F., and B. Knight, 2011, "Media bias and influence: Evidence from newspaper endorsements," *The Review of Economic Studies*, p. rdq037.

Cutler, D. M., J. M. Poterba, and L. H. Summers, 1989, "What Moves Stock Prices?," *Journal of Portfolio Management*, 15, 4–12.

Das, S. R., and M. Y. Chen, 2007, "Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web," *Management Science*, pp. 1375–1388.

DellaVigna, S., and M. Gentzkow, 2010, "Persuasion: Empirical Evidence," *Annual Review of Economics*, 2(1), 643–669.

Dougal, C., J. Engelberg, D. García, and C. Parsons, 2012, "Journalists and the stock market," *Review of Financial Studies*, 25(4), 639–679.

Dyck, A., N. Volchkova, and L. Zingales, 2008, "The corporate governance role of the media: Evidence from Russia," *The Journal of Finance*, 63(3), 1093–1135.

Dyck, A., and L. Zingales, 2003, "The Media and Asset Prices," working paper, University of Chicago.

Ellman, M., and F. Germano, 2009, "What do the papers sell? a model of advertising and media bias," *The Economic Journal*, 119(537), 680–704.

Engelberg, J., 2008, "Costly information processing: evidence from earnings announcements," working paper, University of North Carolina.

Engelberg, J., and C. Parsons, 2011, "The causal impact of media in financial markets," *Journal of Finance*, 66(1), 67–97.

Enikolopov, R., M. Petrova, and E. Zhuravskaya, 2011, "Media and political persuasion: Evidence from Russia," *American Economic Review*, 101(7), 3253–3285.

Fang, L. H., and J. Peress, 2009, "Media coverage and the cross-section of stock returns," *Journal of Finance*, 64(5), 2023–2052.

Foerster, S. R., J. T. Linnainmaa, B. Melzer, and A. Previtero, 2015, "Retail Financial Advice: Does One Size Fit All?," *Journal of Finance*, p. forthcoming.

Gaa, C., 2008, "Good news is no news: asymmetric inattention and the neglected firm effect," working paper, University of British Columbia.

García, D., 2013, "Sentiment during recessions," *Journal of Finance*, 68(3), 1267–1300.

Gentzkow, M., and J. M. Shapiro, 2006, "Media bias and reputation," *Journal of Political Economy*, 114(2), 380–316.

——— , 2008, "Competition and truth in the market for news," *Journal of Economic Perspectives*, 22, 133–150.

——— , 2010, "What drives media slant? Evidence from U.S. daily newspapers," *Econometrica*, 78(1), 35–71.

——— , 2011, "Ideological Segregation Online and Offline," *The Quarterly Journal of Economics*, 126(4), 1799–1839.

Gentzkow, M., J. M. Shapiro, and M. Sinkinson, 2011, "The Effect of Newspaper Entry and Exit on Electoral Politics," *The American Economic Review*, 101(7), 2980–3018.

Griffin, J. M., N. H. Hirschey, and P. J. Kelly, 2011, "How important is the financial media in global markets?," *Review of Financial Studies*, pp. 3941–3992.

Griggs, H., 1963, "Newspaper performance in recession coverage," *Journalism Quarterly*, 40, 559–564.

Groseclose, T., and J. Milyo, 2005, "A measure of media bias," *The Quarterly Journal of Economics*, pp. 1191–1237.

Gurun, U. G., and A. W. Butler, 2012, "Don't believe the hype: local media slant, local advertising, and firm value," *Journal of Finance*, 67, 561–597.

Hanley, K. W., and G. Hoberg, 2012, "Litigation risk, strategic disclosure and the underpricing of initial public offerings," *Journal of Financial Economics*, 103, 235–254.

Heston, S., and N. Sinha, 2016, "News versus sentiment: Comparing textual processing approaches for predicting stock returns," working paper, Working paper, University of Maryland.

Hillert, A., H. Jacobs, and S. Müller, 2014, "Media makes momentum," *Review of Financial Studies*, p. forthcoming.

Huberman, G., and T. Regev, 2001, "Contagious speculation and a cure for cancer: a nonevent that made stock prices soar," *Journal of Finance*, 56, 387–396.

Iyengar, S., and K. S. Hahn, 2009, "Red media, blue media: Evidence of ideological selectivity in media use," *Journal of Communication*, 59(1), 19–39.

Jegadeesh, N., and D. Wu, 2013, "Word power: A new approach for content analysis," *Journal of Financial Economics*, 110, 712–729.

Jurafsky, D., and J. H. Martin, 2018, *Speech and Language Processing*. Prentice Hall, Upper Saddle River, NJ.

Kahneman, D., and A. Tversky, 1979, "Prospect theory: an analysis of decision under risk," *Econometrica*, 47, 263–292.

Klibanoff, P., O. Lamont, and T. Wizman, 1998, "Investor reaction to salient news in closed-end country funds," *Journal of Finance*, 53(2), 673–699.

Kogan, S., B. Routledge, J. Sagi, and N. Smith, 2015, "Information Content of Public Firm Disclosures and the Sarbanes-Oxley Act," working paper, Working paper, UNC Chapel Hill.

Kuhnen, C. M., 2015, "Asymmetric learning from financial information," *The Journal of Finance*, 70(5), 2029–2062.

Larcinese, V., R. Puglisi, and J. M. S. Jr., 2011, "Partisan Bias in Economic News: Evidence on the Agenda-Setting Behavior of U.S. Newspapers," *Journal of Public Economics*, 95(9), 1178–1189.

Loughran, T., and B. McDonald, 2011, "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks," *Journal of Finance*, 66, 35–65.

——— , 2013, "IPO first-day returns, offer price revisions, volatility, and form S-1 language," *Journal of Financial Economics*, 109, 307–326.

——— , 2016, "Textual Analysis in Accounting and Finance: A Survey," *Journal of Accounting Research*, 54, 1187–1230.

Mamaysky, H., and P. Glasserman, 2017, "Does Unusual News Forecast Market Stress?," working paper, Columbia University.

Manela, A., and A. Moreira, 2017, "News Implied Volatility and Disaster Concerns," *Journal of Financial Economics*, 123(1), 137–162.

Mullainathan, S., and A. Shleifer, 2005, "The market for news," *American Economic Review*, pp. 1031–1053.

Neilson, W., 1973, *What's News – Dow Jones*. Clinton Book Company, Radnor, PA.

Neuhierl, A., A. Scherbina, and B. Schlusche, 2013, "Market Reaction to Corporate Press Releases," *Journal of Financial and Quantitative Analysis*, 48(4), 1207–1240.

Norris, F., and C. Bockelmann, 2000, *The New York Times — Century of Business*. McGraw-Hill, New York City, New York.

Peress, J., 2014, "The media and the diffusion of information in financial markets: Evidence from newspaper strikes," *The Journal of Finance*, 69(5), 2007–2043.

Prat, A., and D. Strömberg, 2011, "The Political Economy of Mass Media," working paper 8246, C.E.P.R. Discussion Papers.

Reuter, J., and E. Zitzewitz, 2006, "Do Ads Influence Editors? Advertising and Bias in the Financial Media," *Quarterly Journal of Economics*, 121(1), 197–227.

Roush, C., 2006, *Profits and Losses*. Marion Street Press, Oak Park, Illinois.

Shiller, R. J., 2000, *Irrational Exuberance*. Princeton University Press, Princeton.

Solomon, D., 2012, "Selective publicity and stock prices," *Journal of Finance*, 67(2), 599–637.

Tetlock, P. C., 2007, "Giving content to investor sentiment: the role of media in the stock market," *Journal of Finance*, 62(3), 1139–1168.

——— , 2010, "Does public financial news resolve asymmetric information?," *Review of Financial Studies*, 23(9), 3520–3557.

Tetlock, P. C., M. Saar-Tsechansky, and S. Macskassy, 2008, "More than words: quantifying language to measure firms' fundamentals," *Journal of Finance*, 63(3), 1437–1467.

Williamson, S. H., 2008, "Daily Closing Values of the DJA in the United States, 1885 to Present," working paper, MeasuringWorth.

Table 1
Summary statistics

The table reports sample statistics for the media content measures used in the paper. These measures are constructed from the columns "Financial Markets" and "Topics in Wall-Street" published in the New York Times, as well as "Abreast of the market," from the Wall Street Journal, in the period 1905–2005. We construct the "Positive" and "Negative" measures by counting the number of positive and negative words and normalizing it by the total number of words of each article, using the Loughran and McDonald (2011) dictionaries. The "Media content" variable is simply the difference between the "Positive" and "Negative" measures. All numbers are given in percentages.

	Positive words (%)			Negative words (%)			Media content		
	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
All articles	1.27	1.18	0.63	2.08	1.96	0.94	−0.82	−0.72	1.15
NYT articles	1.21	1.11	0.65	2.08	1.96	0.95	−0.87	−0.79	1.16
WSJ articles	1.42	1.35	0.57	2.09	1.96	0.93	−0.67	−0.54	1.12
Time-series aggregate	1.26	1.23	0.36	2.03	1.97	0.63	−0.77	−0.70	0.79

Table 2
Media content as a function of DJIA returns

The table reports point estimates from the model

$$M_t = \beta_0 R_t + \beta \mathcal{L}(R_t) + \rho \mathcal{L}(M_t) + \eta X_t + \epsilon_t;$$

where M_t denotes the media content written between trading dates t and $t+1$ (for articles written after the market closed on date t by prior to opening on date $t+1$), R_t denotes the log-return on the DJIA on date t (from close on date $t-1$ to close on date t). We truncate R_t at -3% from below and $+3\%$ from above. The set of explanatory variables X_t includes day-of-the-week dummies and a cubic function of time. The t-statistics reported in the table use Newey-West corrections with ten lags.

	Only $\mathcal{L}(R_t)$		Only $\mathcal{L}(M_t)$		$\mathcal{L}(R_t)$ and $\mathcal{L}(M_t)$	
	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat
R_t	0.478	91.7			0.470	95.8
R_{t-1}	0.149	28.5			0.077	13.6
R_{t-2}	0.082	15.7			0.022	3.9
R_{t-3}	0.073	13.9			0.022	3.8
R_{t-4}	0.067	12.8			0.008	1.3
R_{t-5}	0.054	10.3			−0.013	−2.2
R_{t-6}	0.040	7.6			−0.026	−4.6
R_{t-7}	0.049	9.3			−0.002	−0.4
R_{t-8}	0.041	7.9			−0.014	−2.5
R_{t-9}	0.045	8.7			−0.013	−2.4
M_{t-1}			0.217	36.0	0.141	23.3
M_{t-2}			0.071	11.5	0.071	11.7
M_{t-3}			0.058	9.4	0.054	8.8
M_{t-4}			0.068	11.0	0.068	11.2
M_{t-5}			0.056	9.1	0.072	11.8
M_{t-6}			0.045	7.2	0.064	10.5
M_{t-7}			0.023	3.8	0.029	4.8
M_{t-8}			0.039	6.3	0.045	7.5
M_{t-9}			0.036	5.9	0.043	7.1
M_{t-10}			0.044	7.4	0.035	6.7
adj-R^2		0.346		0.224		0.423

Table 3
Media content as a non-linear function of last DJIA return

The table reports point estimates from the model

$$M_t = f(R_t; \alpha, \beta) + \eta X_t + \epsilon_t;$$

where M_t denotes the media content written between trading dates t and $t+1$ (for articles written after the market closed on date t by prior to opening on date $t+1$), R_t denotes the log-return on the DJIA (truncated at -3% from below and $+3\%$ from above). The set of explanatory variables X_t includes 10 lags of R_t and M_t, as well as day-of-the-week dummies and a cubic function of time. The function $f(R_t; \alpha, \beta)$ is assumed to be of the form

$$f(R_t; \alpha, \beta) = \sum_{i=1}^{4} (\alpha_i + \beta_i R_t) 1_{\{R_t \in S_i\}} \tag{6}$$

where the sets S_i are: $S_1 = [-3, -1)$, $S_2 = (-1, 0)$, $S_3 = (0, 1)$, and $S_4 = (1, 3]$. All statistics reported in the table use Newey-West corrections with ten lags. The time period goes from January 3, 1905 through December 31, 2005, for a total of 27,448 trading days.

A. Point estimates	Intercepts		Slopes	
	α_i	t-stat	β_i	t-stat
$S_1 = (-3, -1)$	-0.452	-6.1	0.203	8.6
$S_2 = (-1, 0)$	0.007	0.1	0.656	22.5
$S_3 = (0, 1)$	0.161	2.6	0.528	21.0
$S_4 = (1, 3)$	0.577	7.8	0.057	2.5

B. Tests	F-stat	p-value
$\alpha_2 = \alpha_3$	75.1	0.000
$\beta_2 = \beta_3$	11.3	0.001
$\beta_1 = \beta_4$	20.0	0.000

Table 4
Media content as a function of multiple DJIA lagged returns

The table reports point estimates from the model

$$M_t = f_1(R_t; \alpha_1, \beta_1) + f_2(R_{t-1}; \alpha_2, \beta_2) + f_3(R_{t-2}; \alpha_3, \beta_3) + f_4(R_{t-3}; \alpha_4, \beta_4) + \eta X_t + \epsilon_t;$$

where M_t denotes the media content written between trading dates t and $t+1$ (for articles written after the market closed on date t by prior to opening on date $t+1$), R_t denotes the log-return on the DJIA (truncated at -3% from below and $+3\%$ from above). The set of explanatory variables X_t includes day-of-the-week dummies and a cubic function of time. The function $f_j(R_t; \alpha, \beta)$ is assumed to be of the form

$$f_j(R_{t-j}; \alpha_j, \beta_j) = \sum_{i=1}^{4} (\alpha_{ji} + \beta_{ji} R_{t-i}) 1_{\{R_{t-i} \in S_i\}} \quad (7)$$

where the sets S_i are: $S_1 = [-3, -1)$, $S_2 = (-1, 0)$, $S_3 = (0, 1)$, and $S_4 = (1, 3]$. All statistics reported in the table use Newey-West corrections with ten lags. The time period goes from January 3, 1905 through December 31, 2005, for a total of 27,448 trading days.

A. Point estimates	Intercepts		Slopes	
	α_{ji}	t-stat	β_{ji}	t-stat
1. Return last trading day.				
$S_1 = (-3, -1)$	-0.469	-5.9	0.240	9.3
$S_2 = (-1, 0)$	-0.002	0.0	0.699	22.5
$S_3 = (0, 1)$	0.154	2.3	0.516	19.4
$S_4 = (1, 3)$	0.573	7.4	0.027	1.1
2. Return two-days ago.				
$S_1 = (-3, -1)$	-0.114	-1.5	0.185	7.9
$S_2 = (-1, 0)$	-0.051	-0.8	0.248	8.6
$S_3 = (0, 1)$	-0.053	-0.8	0.109	4.0
$S_4 = (1, 3)$	-0.010	-0.1	0.025	1.0
3. Return three-days ago.				
$S_1 = (-3, -1)$	-0.052	-0.7	0.113	4.7
$S_2 = (-1, 0)$	0.009	0.1	0.155	5.1
$S_3 = (0, 1)$	-0.006	-0.1	0.047	1.7
$S_4 = (1, 3)$	0.092	1.2	-0.057	-2.2
4. Return four-days ago.				
$S_1 = (-3, -1)$	-0.010	-0.1	0.094	3.7
$S_2 = (-1, 0)$	-0.003	-0.1	0.099	3.3
$S_3 = (0, 1)$	0.040	0.6	0.006	0.2
$S_4 = (1, 3)$	0.035	0.4	-0.008	-0.3

Table 4 continued.

B. Tests

	F-stat	p-value
1. Return last trading day.		
$\alpha_2 = \alpha_3$	67.7	0.000
$\beta_2 = \beta_3$	20.5	0.000
$\beta_1 = \beta_4$	35.4	0.000
2. Return two-days ago.		
$\alpha_2 = \alpha_3$	0.1	0.903
$\beta_2 = \beta_3$	12.1	0.001
$\beta_1 = \beta_4$	21.7	0.000
3. Return three-days ago.		
$\alpha_2 = \alpha_3$	0.6	0.444
$\beta_2 = \beta_3$	6.9	0.009
$\beta_1 = \beta_4$	23.7	0.000
4. Return four-days ago.		
$\alpha_2 = \alpha_3$	4.9	0.027
$\beta_2 = \beta_3$	5.4	0.020
$\beta_1 = \beta_4$	8.0	0.005

Table 5
Media content response for different authors

The table reports point estimates from the model

$$M_t = f(R_t; \alpha, \beta) + \eta X_t + \epsilon_t;$$

where M_t denotes the media content written between trading dates t and $t+1$ (for articles written after the market closed on date t by prior to opening on date $t+1$), R_t denotes the log-return on a given stock index (truncated at -3% from below and $+3\%$ from above). The regression is run independently for each of the listed ten authors of the column "Abreast-of-the-market," published in the Wall Street Journal. The column labelled N presents the number of articles for each author. The set of explanatory variables X_t includes 5 lags of R_t and M_t, as well as day-of-the-week dummies and a cubic function of time. The function $f(R_t; \alpha, \beta)$ is assumed to be of the form

$$f(R_t; \alpha, \beta) = \sum_{i=1}^{4} (\alpha_i + \beta_i R_t) 1_{\{R_t \in S_i\}} \tag{8}$$

where the sets S_i are: $S_1 = [-3, -1)$, $S_2 = (-1, 0)$, $S_3 = (0, 1)$, and $S_4 = (1, 3]$. The time period is 1970–2005. All statistics reported in the table use Newey-West corrections with ten lags.

		β_1		β_2		β_3		β_4	
	N	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat
Hillery	2413	0.333	3.3	0.871	9.3	0.636	6.5	0.113	1.3
Obrien	1215	0.533	3.5	1.169	5.1	1.069	5.2	0.218	1.5
Talley	915	−0.025	−0.1	1.111	5.4	1.097	5.3	−0.258	−1.3
Marcial	625	0.899	2.3	0.430	1.5	0.830	2.8	−0.169	−0.5
Garcia	588	0.289	1.4	1.469	6.1	0.597	2.7	−0.112	−0.7
Smith	302	−0.246	−0.6	1.124	3.7	0.945	3.0	−0.078	−0.3
Wilson	251	0.037	0.1	0.988	2.9	0.656	2.4	0.397	1.3
Browning	250	0.592	1.6	0.156	0.4	0.198	0.5	−0.066	−0.2
Pettit	222	0.307	0.4	1.313	3.6	0.218	0.5	−0.587	−0.4
Sease	157	0.624	1.8	0.735	1.6	−0.017	−0.0	0.648	0.8

Table 6
Media content in different time periods

The table reports point estimates from the model

$$M_t = f(R_t; \alpha, \beta) + \eta X_t + \epsilon_t;$$

where M_t denotes the media content written between trading dates t and $t+1$ (for articles written after the market closed on date t by prior to opening on date $t+1$), R_t denotes the log-return on a given stock index (truncated at -3% from below and $+3\%$ from above). The set of explanatory variables X_t includes 5 lags of R_t and M_t, as well as day-of-the-week dummies and a cubic function of time. The function $f(R_t; \alpha, \beta)$ is assumed to be of the form

$$f(R_t; \alpha, \beta) = \sum_{i=1}^{4} (\alpha_i + \beta_i R_t) 1_{\{R_t \in S_i\}} \tag{9}$$

where the sets S_i are: $S_1 = [-3, -1)$, $S_2 = (-1, 0)$, $S_3 = (0, 1)$, and $S_4 = (1, 3]$. Each column presents the point estimates for different subsamples: 1905–1930, 1931–1955, 1956–1980, 1981–2005. All statistics reported in the table use Newey-West corrections with ten lags.

A. Point estimates

	1905–1930		1931–1955		1956–1980		1981–2005	
	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat
β_1	0.241	5.7	0.255	7.3	0.286	4.2	0.236	3.9
β_2	0.459	8.2	0.652	13.0	0.770	14.7	0.794	11.8
β_3	0.375	8.4	0.497	11.8	0.602	12.6	0.604	9.2
β_4	0.024	0.6	0.067	2.1	0.094	1.5	0.116	2.2
α_1	-0.094	-0.8	-0.311	-2.3	-0.716	-4.7	-0.484	-2.3
α_2	0.196	2.1	-0.050	-0.4	-0.219	-1.9	0.122	0.7
α_3	0.319	3.4	0.030	0.3	-0.039	-0.3	0.397	2.1
α_4	0.647	5.5	0.371	2.9	0.355	2.4	0.860	4.2

B. Tests

	F-stat	p-value	F-stat	p-value	F-stat	p-value	F-stat	p-value
$\alpha_2 = \alpha_3$	12.9	0.000	8.4	0.004	28.1	0.000	34.1	0.000
$\beta_2 = \beta_3$	1.3	0.249	5.9	0.015	5.6	0.018	4.1	0.043
$\beta_1 = \beta_4$	12.4	0.000	15.1	0.000	4.2	0.039	2.3	0.129

Table 7
Media content as a non-linear function of other indexes

The table reports point estimates from the model

$$M_t = f(R_t;\alpha,\beta) + \eta X_t + \epsilon_t;$$

where M_t denotes the media content written between trading dates t and $t+1$ (for articles written after the market closed on date t by prior to opening on date $t+1$), R_t denotes the log-return on a given stock index (truncated at -3% from below and $+3\%$ from above). The set of explanatory variables X_t includes 5 lags of R_t and M_t, as well as day-of-the-week dummies and a cubic function of time. The function $f(R_t;\alpha,\beta)$ is assumed to be of the form

$$f(R_t;\alpha,\beta) = \sum_{i=1}^{4}(\alpha_i + \beta_i R_t)1_{\{R_t \in S_i\}} \tag{10}$$

where the sets S_i are: $S_1 = [-3,-1)$, $S_2 = (-1,0)$, $S_3 = (0,1)$, and $S_4 = (1,3]$. Return information is from Ken French's website. In the column labelled "VW index" we use his value-weighted index, whereas in the others we use the largest and smallest quintile portfolios in size. All statistics reported in the table use Newey-West corrections with ten lags. The time period goes from July 1st of 1963 through December 31st of 2005, for a total of 12,958 trading days.

A. Point estimates

	VW index		Large stocks		Small stocks	
	Estimate	t-stat	Estimate	t-stat	Estimate	t-stat
β_1	0.209	4.3	0.281	5.9	0.140	2.5
β_2	0.826	18.6	0.815	18.1	0.874	17.2
β_3	0.682	16.3	0.661	15.5	0.702	15.8
β_4	0.005	0.1	0.055	1.3	0.036	0.6
α_1	−0.580	−4.8	−0.352	−2.8	−0.692	−5.0
α_2	−0.028	−0.3	0.067	0.7	−0.077	−0.7
α_3	0.152	1.7	0.199	2.0	0.084	0.8
α_4	0.685	6.0	0.658	5.5	0.558	3.9

B. Tests

	F-stat	p-value	F-stat	p-value	F-stat	p-value
$\alpha_2 = \alpha_3$	41.2	0.000	20.9	0.000	29.8	0.000
$\beta_2 = \beta_3$	5.7	0.017	6.3	0.012	6.6	0.010
$\beta_1 = \beta_4$	9.2	0.002	12.4	0.000	1.5	0.216

Table 8
Media content and DJIA returns along the business cycle

The table reports point estimates from the model

$$M_t = D_t f_R(R_t; \alpha, \beta) + (1 - D_t) f_E(R_t; \alpha, \beta) + \eta X_t + \epsilon_t;$$

where M_t denotes the media content written between trading dates t and $t+1$ (for articles written after the market closed on date t by prior to opening on date $t+1$), D_t is a dummy variable that equals 1 if date t is in a recession (using NBER definitions), R_t denotes the log-return on the DJIA (truncated at -3% from below and $+3\%$ from above). The set of explanatory variables X_t includes 10 lags of R_t and M_t, as well as day-of-the-week dummies and a cubic function of time. The function $f_k(R_t; \alpha, \beta)$ is assumed to be of the form

$$f_k(R_t; \alpha, \beta) = \sum_{i=1}^{4} (\alpha_{ki} + \beta_{ki} R_t) 1_{\{R_t \in S_i\}} \tag{11}$$

where the sets S_i are: $S_1 = [-3, -1)$, $S_2 = (-1, 0)$, $S_3 = (0, 1)$, and $S_4 = (1, 3]$, and $k = R, E$. All statistics reported in the table use Newey-West corrections with ten lags. The time period goes from January 3, 1905 through December 31, 2005, for a total of 27,448 trading days.

A. Point estimates

	Expansions		Recessions	
	Estimate	t-stat	Estimate	t-stat
β_{k1}	0.236	7.0	0.242	7.4
β_{k2}	0.694	20.6	0.560	9.7
β_{k3}	0.559	19.2	0.404	7.7
β_{k4}	0.050	1.7	0.065	1.8
α_{k1}	−0.503	−5.6	−0.206	−1.4
α_{k2}	−0.021	−0.3	0.092	0.7
α_{k3}	0.139	1.9	0.215	1.7
α_{k4}	0.585	6.7	0.529	3.6

B. Tests

	F-stat	p-value	F-stat	p-value
$\alpha_{k2} = \alpha_{k3}$	60.8	0.000	11.2	0.001
$\beta_{k2} = \beta_{k3}$	9.6	0.002	4.0	0.046
$\beta_{k2} = \beta_{k3}$	16.7	0.000	13.0	0.000

Table 9
Media content as a function of last day's DJIA return, eight intervals

The table reports point estimates from the model

$$M_t = f(R_t; \alpha, \beta) + \eta X_t + \epsilon_t;$$

where M_t denotes the media content written between trading dates t and $t+1$ (for articles written after the market closed on date t by prior to opening on date $t+1$), R_t denotes the log-return on the DJIA (truncated at -3% from below and $+3\%$ from above). The set of explanatory variables X_t includes 10 lags of R_t and M_t, as well as day-of-the-week dummies and a cubic function of time. The function $f(R_t; \alpha, \beta)$ is assumed to be of the form

$$f(R_t; \alpha, \beta) = \sum_{i=1}^{8} (\alpha_i + \beta_i R_t) 1_{\{R_t \in S_i\}} \tag{12}$$

where the sets S_i are: $S_1 = [-3, -2)$, $S_2 = (-2, -1)$, $S_3 = (-1, -0.5)$, $S_4 = (-0.5, 0)$, $S_5 = (0, 0.5)$, $S_6 = (0.5, 1)$, $S_7 = (1, 2)$ and $S_8 = (2, 3]$. All statistics reported in the table use Newey-West corrections with ten lags. The time period goes from January 3, 1905 through December 31, 2005, for a total of 27,448 trading days.

A. Point estimates	Intercepts		Slopes	
	α_i	t-stat	β_i	t-stat
$S_1 = (-3, -2)$	−0.263	−1.2	0.266	3.3
$S_2 = (-2, -1)$	−0.316	−3.0	0.309	4.9
$S_3 = (-1, -0.5)$	−0.142	−1.5	0.455	4.7
$S_4 = (-0.5, 0)$	0.021	0.3	0.711	11.3
$S_5 = (0, 0.5)$	0.102	1.6	0.806	13.3
$S_6 = (0.5, 1)$	0.280	3.3	0.346	4.4
$S_7 = (1, 2)$	0.543	5.4	0.079	1.4
$S_8 = (2, 3)$	0.996	4.8	−0.100	−1.3

B. Tests	F-stat	p-value
$\alpha_4 = \alpha_5$	11.6	0.001
$\beta_4 = \beta_5$	1.2	0.271
$\beta_3 = \beta_6$	0.7	0.387
$\beta_2 = \beta_7$	7.5	0.006
$\beta_1 = \beta_8$	11.3	0.001

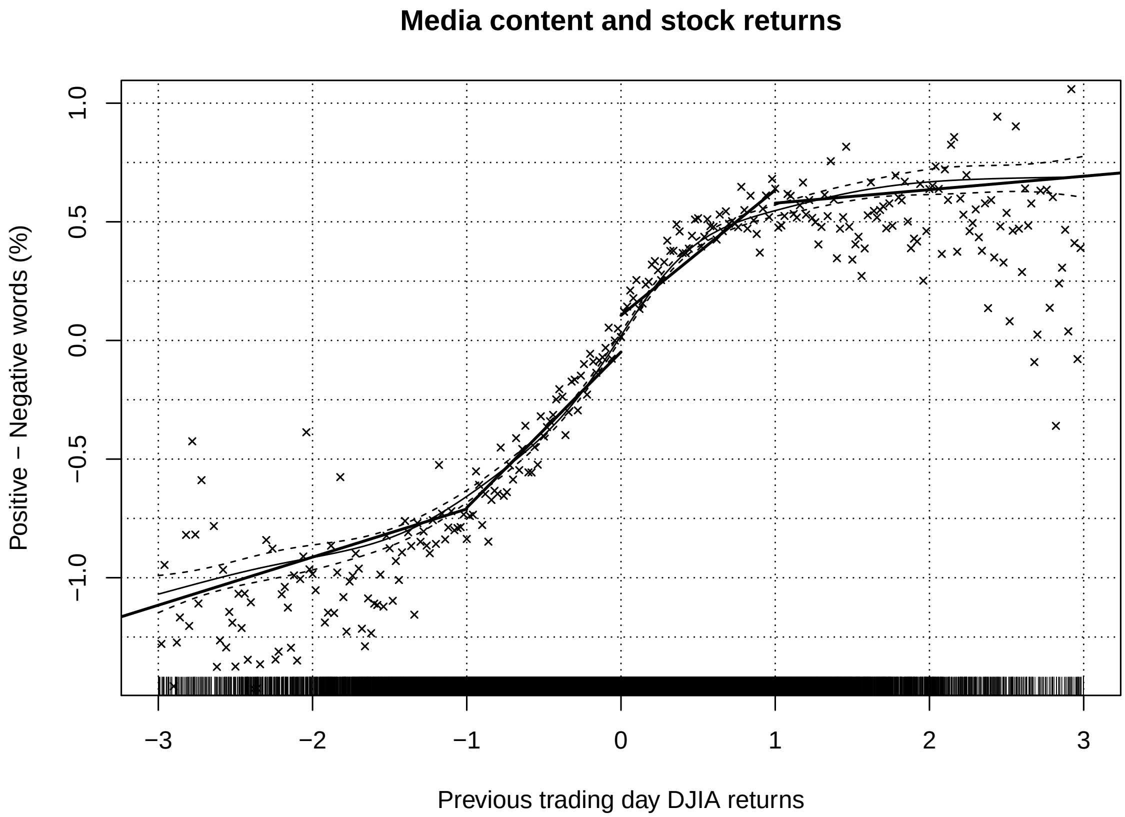

Figure 1
The graph presents three estimates of the relationship between media content and lagged DJIA returns. The crosses denote the average unconditional media content for intervals of 2 basis points on the range of returns $(-3, 3)$. The solid smooth black line presents a spline estimator, with 95% confidence intervals as dashed lines. The straight lines are OLS estimates using piece-wise linear functions over the intervals $(-3, -1)$, $(-1, 0)$, $(0, 1)$ and $(1, 3)$. The rug in the x-axis presents the density of the right-hand side variable.

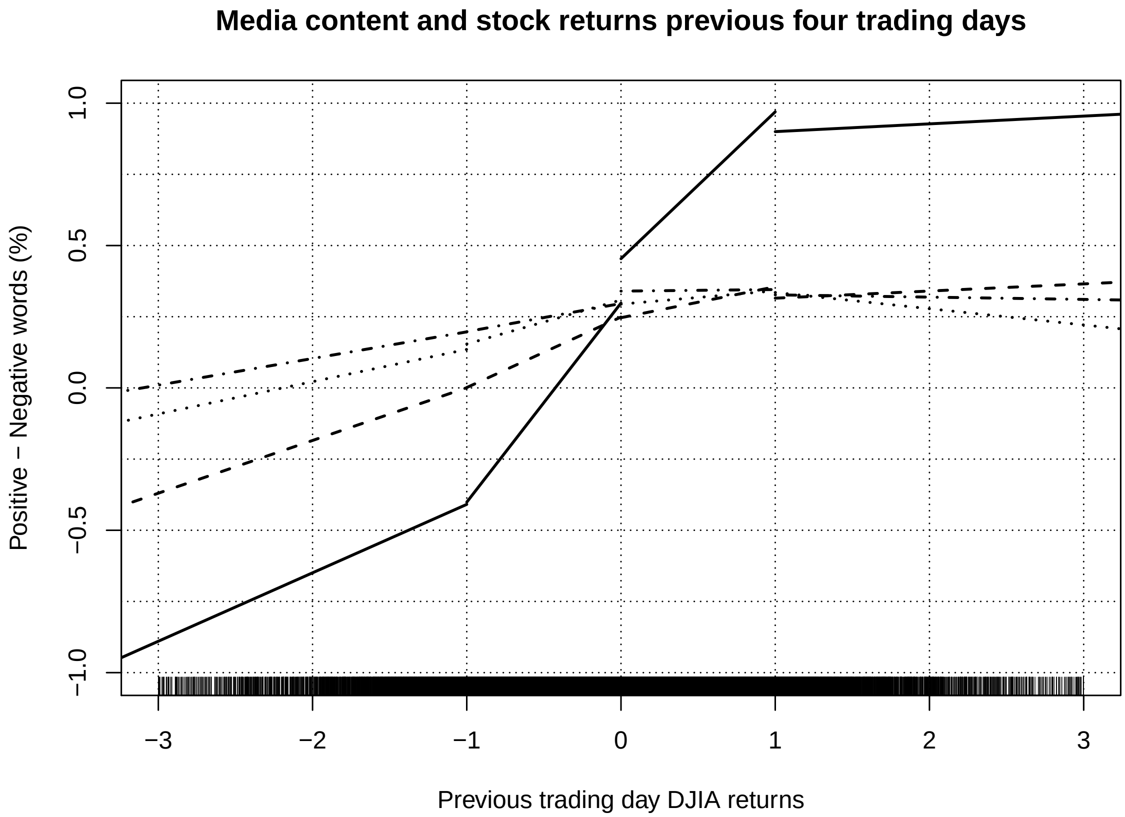

Figure 2
The graph presents four estimates. They measure the relationship between media content and: (1) one-day lagged DJIA returns (solid line), (2) two-day lagged DJIA returns (dashed line), (3) three-day lagged DJIA returns, and (4) four-day lagged DJIA returns. The lines in the graphs represent the OLS estimates using piece-wise linear functions over the intervals $(-3,-1)$, $(-1,0)$, $(0,1)$ and $(1,3)$ of the specification in Table 4. The rug in the x-axis presents the density of the right-hand side variable.

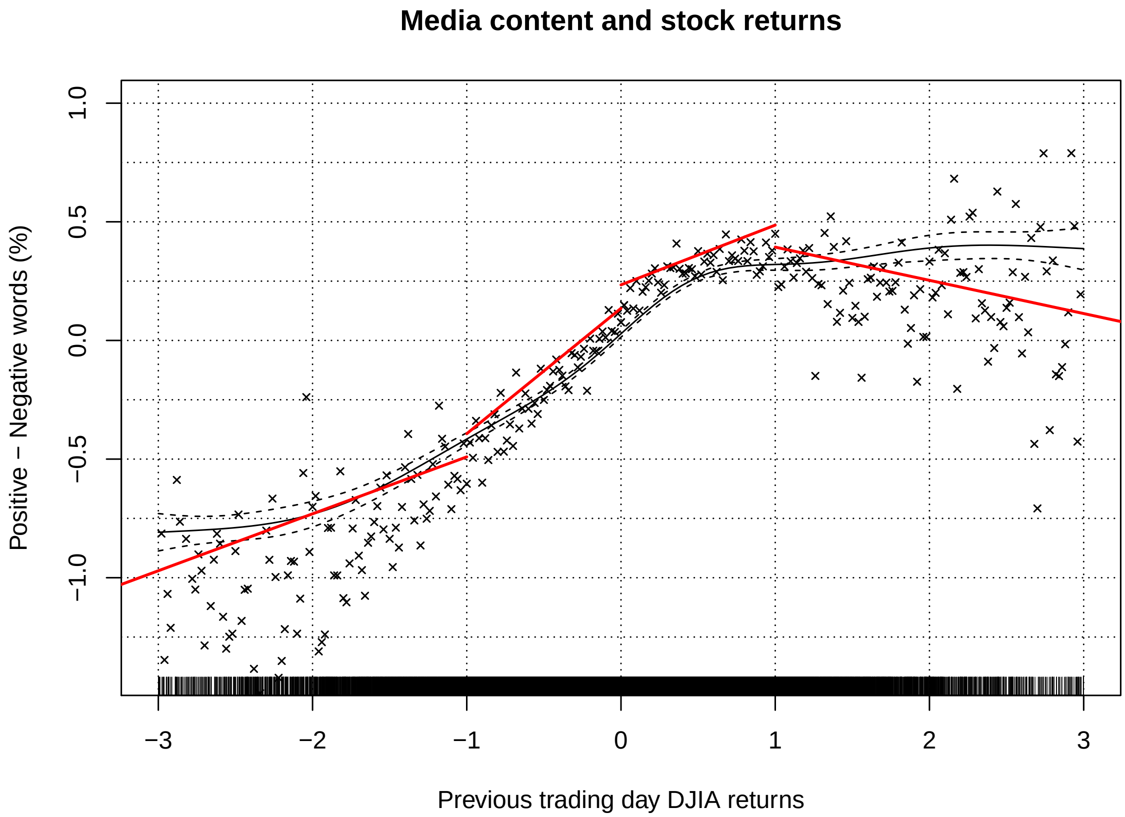

Figure 3
The graph presents three estimates of the relationship between media content and lagged DJIA returns. Media content is measured using `tf-idf` weights, which emphasize words that are used less often. The crosses denote the average unconditional media content for intervals of 2 basis points on the range of returns $(-3, 3)$. The solid smooth black line presents a spline estimator, with 95% confidence intervals as dashed lines. The straight lines are OLS estimates using piece-wise linear functions over the intervals $(-3, -1)$, $(-1, 0)$, $(0, 1)$ and $(1, 3)$. The rug in the x-axis presents the density of the right-hand side variable.

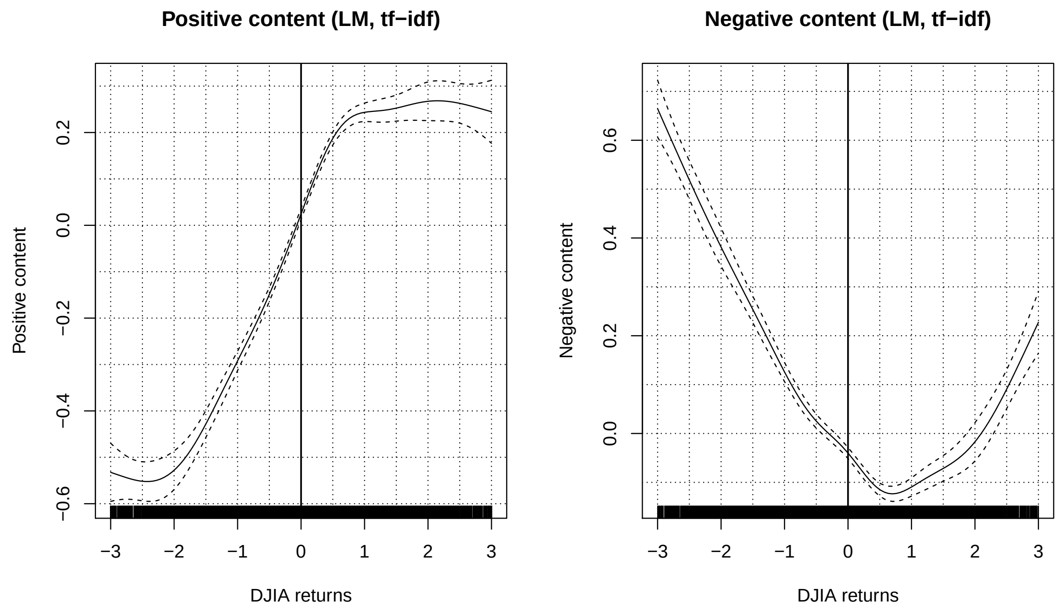

Figure 4
The two graphs present a spline estimator, with 95% confidence bands as dashed lines, of the relationship between media content and lagged DJIA returns. Media content is measured using `tf-idf` weights of positive words (left-panel) and negative words (right-panel). The rug in the x-axis presents the density of the right-hand side variable.